The Dissertation Committee for Gabriel Chi Sun Wu
certifies that this is the approved version of the following dissertation:

Elucidation of the Human B Cell Immune Repertoire by High-Throughput Sequencing and Computational Simulation

Committee:

Edward M. Marcotte, Supervisor

Claus O. Wilke

George Georgiou

Haley Tucker

Michael J. Daniels

Elucidation of the Human B Cell Immune Repertoire by High-Throughput Sequencing and Computational Simulation

by

Gabriel Chi Sun Wu, B.S.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

December 2016

For those who dream bigger than they should,
they achieve what others believe impossible.

Acknowledgments

It has been a long and at times difficult journey. I could not have made it this far without the staggering generosity, patience, and support from my advisor, Edward Marcotte. I am heavily indebted to him.

The caliber of faculty who helped me in this process has been an embarrassment of riches. Sara Sawyer championed me early in my graduate career and I owe much of my early success to her. Jeffrey Barrick was highly supportive and helpful in my first year of graduate school. My committee (Claus Wilke, George Georgiou, Haley Tucker, and Michael Daniels) has been cheerful and engaged. Gregory Ippolito has been an amazing teacher.

The number of people it takes to keep a graduate student well supplied and connected is immense. I am grateful to the staff members who helped take care of enough things so that I could focus on my research. Ashley, Brandy, Jorge, Sean, Lizzie, Jolie, Tom, Jen G, Brooke, Barbara, Jennifer W, and last but certainly not least Rob, we do not give you all enough credit for keeping the the lights on, the computers running, and the supplies stocked.

My past and present colleagues: JJ, Kam, Costa, Ellen, Will, Sebastian, and Bing. I hesitate to call myself an immunologist, but if I am ever mistaken for one, it is because of what you have taught me.

My lab is a place of never ending activity, excitement, and distraction. Peggy and Martin, you were the first to greet me and welcome me to the lab.

Blake, your focus and clarity of thought are inspiring. John H and Kevin D, your aptitude and wit provide direction and enjoyment to the lab. Ophelia and Chris, you bring a little bit of sanity to the lab–not too much, of course. Dan and Andrew, our stimulating conversations were the best if not the most productive times. Hyeji, Taejoon, Angela, Jeremy, Jag, Jon L, Aashiq, Claire, Alex, Alice, Mark, Anna M, Matt, and Ben L you have all been wonderful co-conspirators of ideas and amusement. I adore all of you and the rest of my labmates for building a truly exceptional workplace in both competence and character.

Jon and Ben, you are true friends. Thank you.

My friends and mentors before my life in Austin, especially the Normal Weirdos, JAGS, Dueber Lab, Anderson Lab, Howard Maibach, Raja Sivamani, Ian Holmes, Jay Keasling, and Adam Arkin. You gave me the confidence to pursue graduate studies and have continued your support from afar.

Anna, I am here today because of you. Our many adventures and endless conversation challenge and delight me. Your expectations and aspirations for me exceed those that I have for myself; I try everyday to be worthy of such devotion.

Finally, I cannot express how thankful I am for the constant support of my family. To my parents and my brother, your unconditional love and relentless loyalty is awe inspiring and humbling.

Thank you all.

Elucidation of the Human B Cell Immune Repertoire by High-Throughput Sequencing and Computational Simulation

Publication No. _____

Gabriel Chi Sun Wu, Ph.D.
The University of Texas at Austin, 2016

Supervisor: Edward M. Marcotte

The human immune system carefully balances the need to maintain stable responses to familiar stimuli with the need for agile responses to an ever changing array of potential dangers. Classic techniques allow for detailed evaluation of parts of the immune system, while emerging technologies allow for more systems-level analysis of the immune system as a whole. In this dissertation, I use high-throughput techniques and computational analysis to advance our understanding of the human bone marrow B cell repertoire. First, I describe the variation in composition of human bone marrow plasma cells from the same individual over time. I show that the frequency of gene and gene combination usage, assayed by high-throughput sequencing, is temporally stable over 6.5 years. Next, I describe a computational model that simulates the process of high-throughput sequencing of immune cells and identify the major

sources of error in these experiments. Specifically, this simulation demonstrates that the typical shape of the experimental distribution of antibodies may be in large part be due to error generated in the experimental process and not a biologically relevant observation. I go on to demonstrate the current limits in understanding the initial distribution of the immune repertoire due to accumulated noise in the experimental process. The work presented here represents the longest longitudinal study to date of high-throughput sequencing techniques used to study the repertoire of human B cells. In addition, the computational model frames the technical challenges of immunological repertoire analysis. This knowledge will provide the basis of future studies to understand the nature of B cells in human bone marrow. It will be relevant for both academic and clinical researchers studying the immune system at basal state as well as at an active defense state. Ultimately, it provides guidance to the community at large with the intent of improving immunology and human health.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The immune system is unique in its potential for diversity. The need for diversity stems from its role in preventing disease progression in all its forms [61]. The diversity of the immune system manifests itself on many levels, including at the level of cellular compartment, cell type, or cell receptor. At the level of the B cell receptor, or antibody, this diversity is accomplished by genomic rearrangement, junctional diversity, and somatic hypermutation (SHM) [25]. In humans, estimating the theoretical limit of diversity using genomic rearrangement alone results in 2.3×10^6 antibody arrangements [20]. Junctional diversity increases the possible antibody sequences to 10^{11} [20]. SHM increases the diversity even more. Despite this extraordinarily large amount of possible diversity, the actual available repertoire has been estimated to be 10^7 [5]. Nevertheless, the true size and degree of diversity of the antibody repertoire remains unknown. Additional measurements of the size, diversity, and temporal dynamics of the immune repertoire will provide knowledge of a person's environmental history and innate ability to handle any future encounters with disease causing agents.

There are several major cellular reservoirs of antibody producing cells

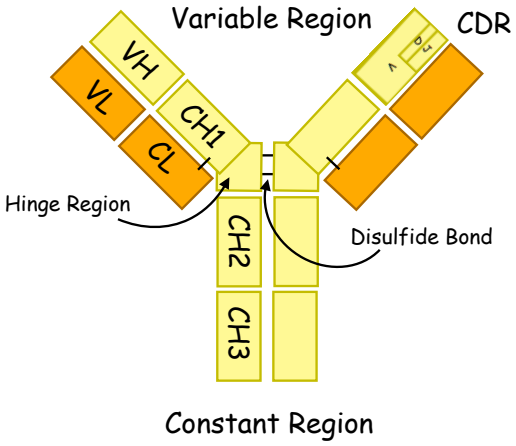


Figure 1.1: **The structure of IgG**

in humans: the blood, bone marrow, lymph nodes, thymus, and spleen among others. The most easily accessible (and heavily researched) of these compartments is the blood, where B cells in the peripheral blood mononuclear cells (PBMCs) are readily accessible. However, the major reservoir of antibody producing cells is in the bone marrow, where antibody producing cells can survive for long periods of time [45]. The actual lifetime of these antibody producing cells remains an open and important question. The answer would reveal how long our immune system can effectively maintain an active response to past stimuli.

The structure of the antibody balances the need for effective defense and enormous diversity. Antibodies are homodimeric tetramers (Figure 1.1). The dimer is made of two polypeptide chains referred to as heavy and light chains. The heavy and light chains are composed of a variable and constant region. The variable heavy chain is composed of three genes (V, D, and J).

The light variable chain is composed of two genes (V and J). There are five classes (IgM, IgD, IgG, IgA, and IgE) of heavy chains and two classes of light chains (lambda and kappa). The mature molecule is a 150 kDa protein molecule. The separate variable light and heavy chains and the size of each molecule make full length sequencing challenging. However, there are regions in the variable regions that are hypervariable [68]. Particularly, the region encompassing the end of the V gene, the entire D gene, and the beginning of the J gene, known as the CDR-H3. This region is typically short enough to be covered by sequencing reads; therefore, by sequencing the variable heavy chain, a measure of antibody diversity and clonality can be calculated.

In recent years, the development of high-throughput sequencing (HTS) has enabled researchers to interrogate the immune system at a previously unattainable depth [46]. HTS technologies have been used to find biases in antibody gene selection and rearrangements across patients [4, 25]. They have been used to look at antigen induced responses, specifically against HIV-1 [65] and influenza [33, 39, 36, 63]. In addition, they have been important in technologies that identify heavy and light chain pairing [14, 9].

There are two primary sequencing technologies used for antibody repertoire analysis [26]. The first, 454 GS FLX Titanium is based on a method of sequencing by synthesis using single-nucleotide addition. While having the benefits of relatively long reads (600-1000bps), it suffers from susceptibility to homopolymer misidentifications, high error rate ($\sim$1%), and high cost per run. The second and more prevalently used method is Illumina MiSeq. This

technology is based on sequencing by synthesis using cyclic reversible termination. The maximum read length is shorter than 454 GS FLX Titanium at 600 bps (achieved by 2×300 paired end reads), but it has a lower error rate (0.1%) and cost per run. No matter what platform is used, there are many challenges of HTS that are prominent specifically for antibody repertoire sequencing. First, the antibody repertoire is highly variable across individuals preventing the generation of a reference sequence. In addition, because antibodies are the product of multiple gene recombination, junctional diversity, and short hypervariable regions separated by long highly similar sequences, short reads are insufficient to identify antibody sequences accurately.

The fundamental question on how best to describe antibody diversity remains open. To better understand these intricacies, the antibody repertoire community has looked to ecology for various methods to measure diversity [43]. The total population of immune cells is much like an ecosystem. In addition, the difficulty of sampling an immune system is similar to that of sampling an entire ecosystem [5]. The methods of sampling, therefore, are highly analogous. Many different metrics of diversity exist, e.g. Parker-Berger, Shannon, and Simpson. These metrics tend to focus on two characteristics: species richness and species evenness, generally emphasizing one trait more than the other. Recently, specific indices have been suggested for immune repertoire diversity [35]. Diversity index choice is highly context dependent; therefore, the assumptions and biases of those diversity indices should be defined when used for biological interpretation. A well controlled and repeatable framework

for that analysis remains absent for the community.

In this dissertation, I explore the immune repertoire at the level of the bone marrow plasma cells and further interrogate the methods of immune repertoire analysis via simulation. In Chapter 2, I observe the longevity of antibody producing cells. I analyzed antibody producing cells across several timepoints from the same human donor and explored the temporal dynamics of gene usage and rearrangement, CDR-H3 length, and isotype usage. In addition, by analyzing the B cell bone marrow repertoire, I take one step toward furthering our understanding of the human immune response. In Chapter 3, I build a simulation of error modes in order to understand the extent to which the HTS empirical immune repertoire can be used to interpret the true biological immune repertoire. I demonstrate that there are multiple sources of error, including cell sampling, nucleic acid amplification, and high-throughput sequencing. I quantify the extent of error correction that can be achieved by sequence clustering. Surprisingly, I find that different true biological distributions, when processed through our simulation of HTS immune repertoire workflow, result in the same empirical distribution. Based on these results, I provide concrete recommendations to others in the field for performing immune repertoire studies.

Chapter 2

Temporal Stability and Molecular Persistence of the Bone Marrow Plasma Cell Antibody Repertoire[1]

2.1 Abstract

Plasma cells in human bone marrow (BM) are thought to be responsible for sustaining lifelong immunity, but its underlying basis is controversial. Using high-throughput sequence analysis of the same individual across 6.5 years, we show that the BM plasma cell immunoglobulin heavy chain repertoire is remarkably stable over time. We find a nearly static bias in individual and combinatorial gene usage across time. Analysis of a second donor corroborates these observations. We also report the persistence of numerous BM plasma cell clonotypes ($\sim$2%) identifiable at all points assayed across 6.5 years, supporting a model of serological memory based upon intrinsic longevity of human plasma cells. Donors were adolescents who completely recovered from neuroblastoma prior to the start of this study. Our work will facilitate differentiation between

[1]Published as Wu GC, Cheung NV, Georgiou G, Marcotte EM, Ippolito GC, Temporal Stability and Molecular Persistence of the Bone Marrow Plasma Cell Antibody Repertoire. *bioRxiv*, (2016): 066878. Also accepted for publication at Nature Communications. GG and NKC conceived the study. GCW and GCI designed and performed experiments, analyzed data, prepared figures, and wrote the manuscript, under the supervision of EMM. All authors reviewed the manuscript.

healthy and diseased antibody repertoires, by serving as a point of comparison with future deep sequencing studies involving immune intervention.

2.2 Introduction

The human bone marrow (BM) is a specialized immune compartment that is responsible for both the initial generation of newly formed B cells and the maintenance of terminally differentiated, antibody-secreting plasma cells. The BM, and the plasma cells it harbors, is a central site of antibody production and is the major source of all classes and subclasses of human immunoglobulins (Ig) in the serum [47, 6]. Ig-secreting BM plasma cells are generally believed to be long-lived and to persist for the lifespan of the organism [51]. Longitudinal serological studies have established that antiviral serum antibodies can be remarkably stable, with half-lives ranging from 50 years (e.g. varicella-zoster virus) to 200 years (e.g. measles and mumps); however, by contrast, antibody responses to non-replicating antigens (e.g. tetanus and diphtheria bacterial toxins) rapidly decay with much shorter half-lives of only 10-20 years [2]. These differences not only suggest that antigen-specific mechanisms have a substantial role in the establishment and/or maintenance of serological memory, but raises the question of whether the differential stability of antibody responses might reflect differential intrinsic longevity of plasma cells. This mechanism has been previously proposed in the context of vaccinations and infections [2, 28], and is also supported by observations of differential stability of autoantibody titers when using B cell depleting therapies to treat

autoimmune diseases [11, 10].

The basis of lifelong serological memory (antibody responses) is controversial [51, 3, 69]. A model for intrinsic longevity in plasma cell survival (and hence longevity in serum antibody maintenance) has been posited for the laboratory mouse [44, 54], but data for human plasma cells have not been generated. On the basis of mouse models, human BM plasma cells are assumed to be similarly long-lived and the major source of serum antibodies; however, the contribution of antigen-specific BM plasma cells in humans has only recently been shown experimentally [28, 48]. Despite these notable advances, the availability of corresponding molecular data (namely, sequence data of BM plasma cell Ig transcripts) and of information regarding plasma cell dynamics *in vivo* is scarce. Persistent antigens as well as the memory B cell compartment are implicated in alternative models of lifelong serological memory, implying continual clonal replacement of antigen-specific plasma cells, in contrast to intrinsic plasma cell longevity [45, 7, 55].

Three studies have generated BM plasma cell data using next-generation sequencing techniques, but did not examine the temporal changes that occur in the antibody repertoire over time [28, 13, 57]. Here, building upon our prior experiences with the comprehensive analysis of human cellular and serological antibody repertoires [14, 31, 62, 40, 15], we present the first longitudinal study of serially acquired human BM plasma cells assayed by next-generation deep sequencing. To directly measure the temporal dynamics of BM plasma cells—and to indirectly gain insight into long-lived serological memory—we sequence

recombined VHDJH regions (cDNA), which encode the variable domain (protein) of antibody IGH heavy chains. Most of the VHDJH genetic diversity is in the CDR-H3 hypervariable interval (encoded by a D element, random non-templated nucleotides, and small portions of the VH and JH elements). CDR-H3 is a primary determinant of antibody specificity [66, 32] and has long been considered a unique fingerprint which aids identification of a progenitor B cell and its clonal progeny (B cell clonotype) [67]. We sequence BM plasma cells from the same individual at seven time points over a total of 6.5 years and from a second individual with two timepoints over 2.3 years. The temporal resolution and duration of sampling provides a method to interrogate the *in vivo* temporal dynamics of BM plasma cells in a previously uncharacterized way. We provide detailed temporal information on the individual genes (IGH V, D, and J), gene combinations (V-D, V-J, D-J, V-D-J), and temporally persistent CDR-H3 clonotypes. The second individual provides support that our observations are not unique. Moreover, persisting CDR-H3 clonotypes are class-switched and somatically mutated (in the IGHV gene segment) implying derivation from activated B cell progenitors that must have been selected by antigen. Crucially, persisting CDR-H3 clonotypes reside exclusively in the plasma cell compartment, but are absent among comparable memory B cells (also a class-switched and somatically mutated B cell compartment) isolated from the same BM biopsy. Overall, our results underscore the temporal stability of the IGH V region repertoire according to multiple metrics (temporally stable IGH molecular phenotypes), and provide unequivocal sequence-based

evidence for the persistence of plasma cell clonotypes spanning 6.5 years.

2.3 Results

2.3.1 High-throughput sequencing of serial bone marrow biopsies

To investigate the temporal dynamics of the IGH antibody gene repertoire of bone marrow (BM) plasma cells, we sampled, sorted, and performed high-throughput sequencing (Figure 2.1a). Serial bone marrow biopsies were obtained from two adolescents (Table 2.1) as part of routine evaluations for non-immuno-hematological disease. BM plasma cells were isolated using FACS (fluorescence-activated cell sorting) for CD38++ CD138+ cells within the mononuclear light-scatter gate (Figure 2.1b). Additionally, the cells were uniformly positive for the TNF-receptor superfamily member CD27 (Figure 2.1b, inset). Importantly, we avoided gating of the pan-B cell marker CD19 since previous characterizations of human BM plasma cells show heterogeneous expression of CD19 [34, 21]. Therefore, our method captured all recently described BM plasma cell subpopulations [28, 48] with an overall CD19+/- CD27+ CD38++ CD138+ phenotype. Subsequently, transcripts were amplified from BM plasma cells expressing IgM, IgG, and IgA using RT-PCR followed by high-throughput sequencing.

In total, 503,415 total sequencing reads were generated from 51,200 BM plasma cells (see Methods and Table 2.1). These data span seven timepoints across 6.5 years for Donor 1 (Figure 2.1c) and two timepoints across 2.3 years for Donor 2. A biological replicate (i.e., a second frozen ampule derived from

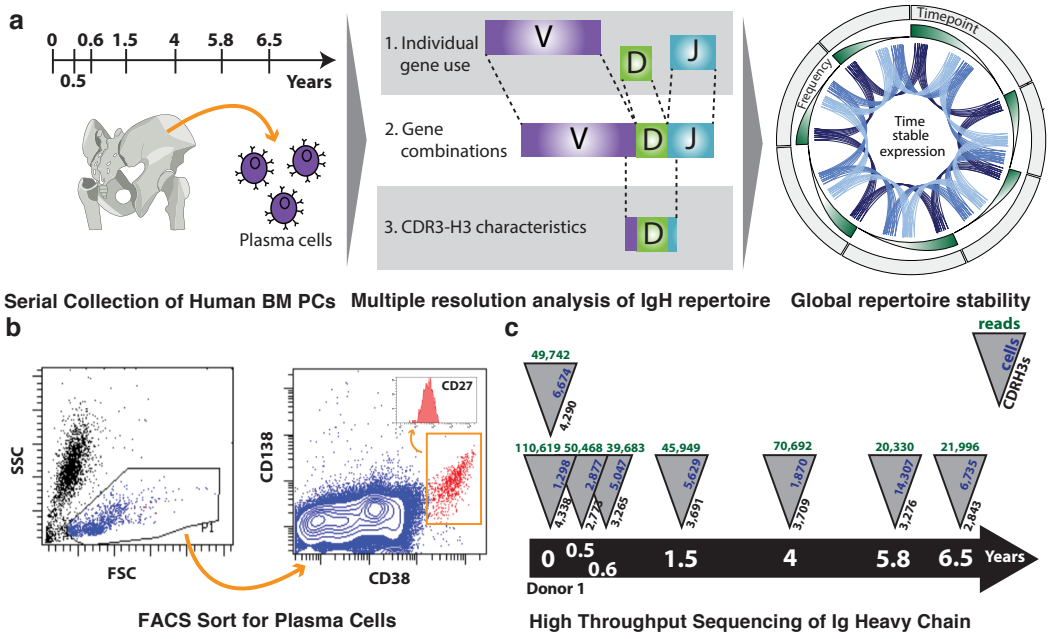


Figure 2.1: **Overview of bone marrow plasma cell sampling and NGS** (a) Overview of antibody repertoire characterization method. Serial sampling of human bone marrow (BM) plasma cells over 6.5 years (left). Analysis of individual genes, gene combinations, and CDR-H3s (center) show temporally stable expression of persistent entities (right). (b) Representative fluorescence-activated cell sorting (FACS) gates of BM plasma cells (CD138+, CD38++) isolated from bone marrow mononuclear cells (BMMCs). (c) Sample collection timeline and summary of cell counts, quality-filtered sequencing reads, and unique CDR-H3s for Donor 1.

Sample ID	Donor	Age (years)	Time (years)	Cells counted	Read counts	Unique CDR-H3s
d1t00a	1	10.9	0	6,674	49,742	4,290
d1t00b	1	10.9	0	1,298	110,619	4,338
d1t05	1	11.2	0.5	2,877	50,468	2,773
d1t06	1	11.5	0.6	5,047	39,683	3,265
d1t15	1	12.4	1.5	5,629	45,949	3,691
d1t40	1	14.9	4	1,870	70,692	3,709
d1t58	1	16.5	5.8	14,307	20,330	3,276
d1t65	1	17.3	6.5	6,735	21,996	2,843
d2t00	2	13.5	0	3,642	17,726	2,120
d2t23a	2	15.78	2.28	2,021	39,096	2,855
d2t23b	2	15.78	2.28	1,100	37,114	4,999
Total				51,200	503,415	38,159

Table 2.1: **Donor history and sequencing information** Bone marrow (BM) plasma cells were isolated from each sample by flow cytometry. BM plasma cells are defined as CD138+ CD38++ cells from bone marrow mononuclear cells. See Figure 2.1 and Methods. Donor 1 was diagnosed at the age of 9 years with adrenal neuroblastoma metastatic to the bone marrow. Donor underwent multiagent chemotherapy consisting of high dose alkylators, then consolidated with myeloablative therapy followed by hematopoietic stem cell transplant. Because of progressive disease in bone marrow and bones at age 10, local radiation and systemic 131 I-MIBG was given followed by anti-GD2 antibody immunotherapy, 3F8+ GM-CSF+ beta-glucan+ 13-cis- retinoic acid till age 14. Donor continued in remission through age 17 years. Because of cancer therapy, donor had to be re-immunized with tetanus, Hemophilus influenza b (Hib), Hepatitis B, and Polio at age 12 (before sample d1t15) and boosted again with Hib, Hepatitis B and Polio at age 13 (between sample d1t15 and d1t40). MMR (mumps measles rubella) vaccine was then given at age 14 (before sample d1t40 and d1t58 and d1t65). Donor 2 was diagnosed at the age of 4 with mediastinal neuroblastoma metastatic to bone and bone marrow and received high dose multiagent chemotherapy. Tumor recurred as epidural mass in the lumbar at the age of 12 and was retreated with high dose multiagent chemotherapy followed by myeloablative therapy plus autologous hematopoietic stem cell rescue and focal radiation to the spine. Donor was treated with anti-GD2 3F8 immunotherapy plus oral etoposide till age 14, and remained in remission through age 20 years.

the same bone marrow aspiration) was also collected from each donor and analyzed. Replicate sampling from the same donor and time point allowed us to confidently discern the active spectrum of heavy chain genes comprising a donors antibody repertoire.

2.3.2 Individual gene frequencies are highly stable

For Donor 1, we identified 38 IGHV genes, 21 IGHD genes, and 6 IGHJ genes (4,788 combinations). We assessed the frequency of each IGH V, D, and J gene across time (Figure 2.2) and found stability of individual gene usage. The most frequently used genes (e.g. IGHV4-34) show consistently high expression while less frequently used genes (e.g. IGHV3-72) show consistently low expression. This observation was quantified using the Mann-Kendall Test, which evaluates trends in time series data. We find that 89% of IGHV genes, 95% of IGHD genes and 100% IGHJ genes show no statistically significant trends (Mann-Kendall test, $p>0.05$), indicating that the IGHV (Figure 2.2a), IGHD (Figure 2.2b), and IGHJ (Figure 2.2c) genes are time stable.

Next, we analyzed population behavior of gene usage. Averaging across all timepoints, we observe a highly skewed distribution of individual gene frequencies, consistent with previous single timepoint observations. Only 6 IGHV genes (16%) account for greater than 50% of total IGHV gene usage by frequency (Figure 2.2d). IGHD2-2, IGHD3-3, and IGHD3-22 [38], previously shown to have biased usage, together account for 33% of total IGHD usage (Figure 2.2b). In addition, known biases in IGHJ usage [59] are recapitulated.

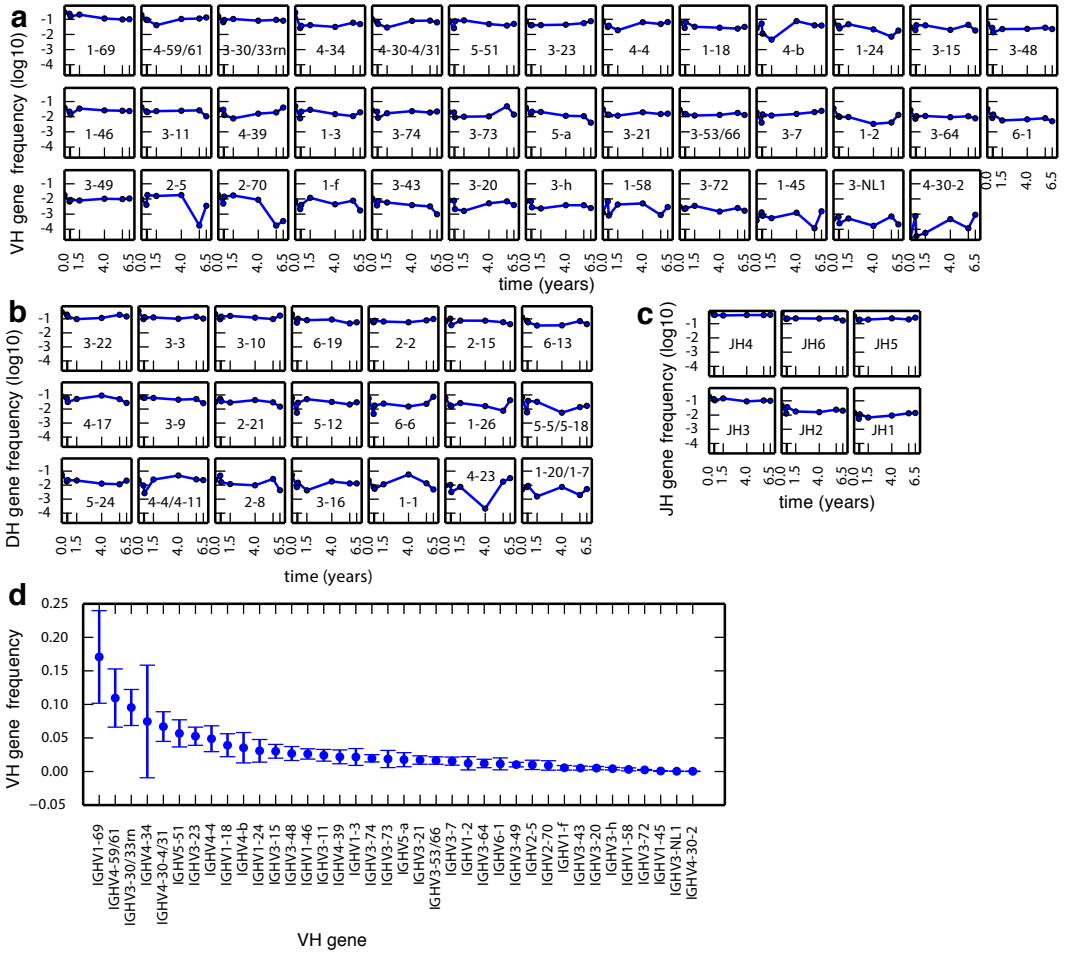


Figure 2.2: **IGH gene segment frequencies among BM plasma cells are temporally stable** For Donor 1: (a-c) IGHV (a), IGHD (b), and IGHJ (c) gene usage frequency over time. Plots are sorted by decreasing mean frequency. Only gene identifications that appear in all timepoints are shown. (d) Mean frequency of IGHV gene use. Error bars are standard deviation.

IGHJ4, IGHJ6, and IGHJ5 account for 86% of total IGHJ usage. Furthermore, IGH V, D, and J gene usage are not significantly different from a log-normal distribution (Anderson-Darling, H=0, p>0.05).

2.3.3 Gene combination frequencies are stable over time

Given the temporal stability of individual genes, we hypothesized that differential intrinsic longevity might be found in gene combinations. Surprisingly, our analysis indicates that gene combinations, like their individual component genes, are time stable as well. We find that 92% V-J (Figure 2.3), 97% V-D (Figure 2.4), 95% D-J (Figure 2.5), and 97% V-D-J (Figure 2.6) do not show significant trends (Mann-Kendall, H=0, p>0.05).

To better understand the nature of gene combinations, we analyzed preferential gene pairing biases by comparing the expected versus observed frequency of pairwise gene combinations. The observed frequency of each gene combination is correlated to its expected frequency (Spearman r): V-D (0.74), V-J (0.87), D-J (0.93), and V-D-J (0.65) (Figure 2.7a-d). This high level of correlation and lack of significant outliers suggests minimal gene pairing linkage and that gene pairing is a random process.

2.3.4 Persistent CDR-H3 clonotypes are unique to BM plasma cells

To understand how each of these individual genes and gene combinations together might indicate the existence of long lived plasma cells, we analyzed the behavior of the CDR-H3, the highest resolution possible for a single

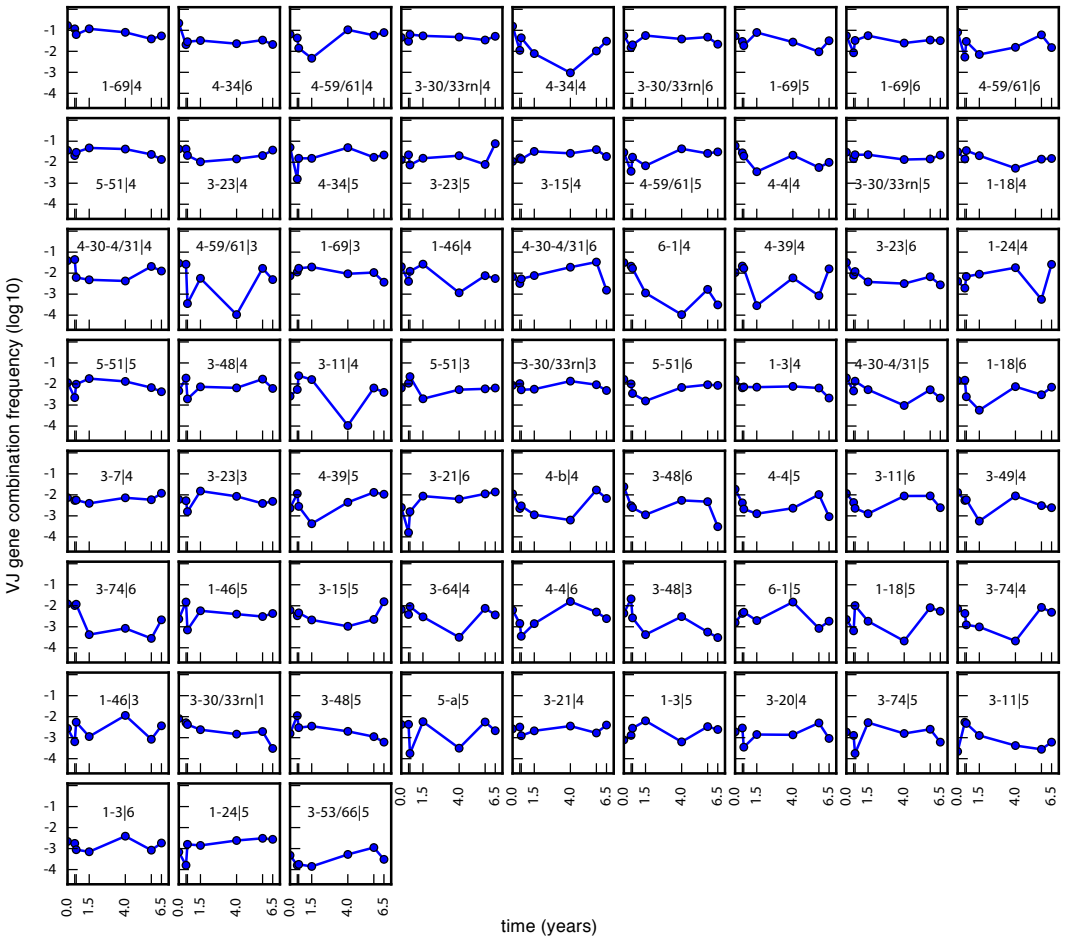


Figure 2.3: **Frequencies of gene combinations among BM plasma cells are temporally stable** IGH V-J usage frequencies for Donor 1 are shown. Plots are sorted by decreasing mean frequency. Only gene identifications that appear in all timepoints are shown. See Figure 2.4(a-c) for usage frequencies of IGH V-D, D-J, and V-D-J.

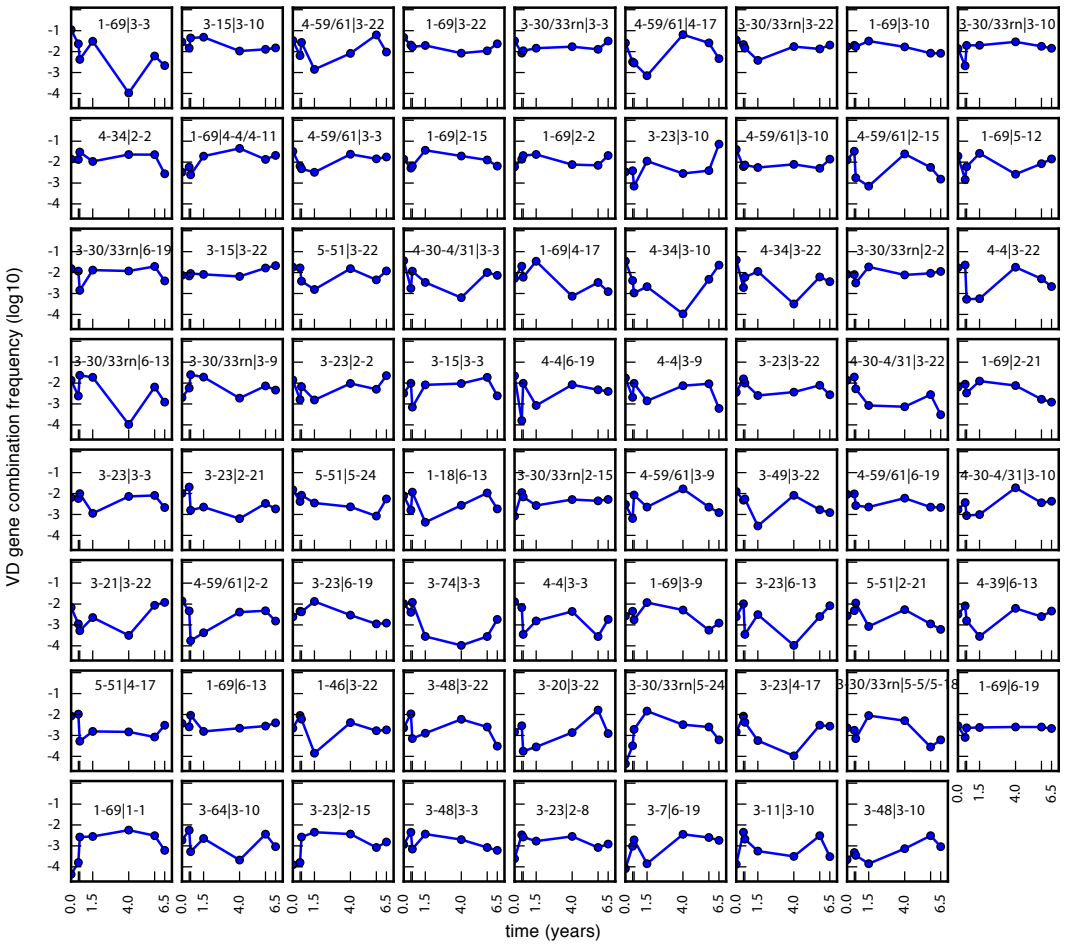


Figure 2.4: **IGH V-D combination gene use frequency of plasma cells from Donor 1** Plots are sorted by decreasing mean frequency. Only gene identifications that appear in all timepoints are shown.

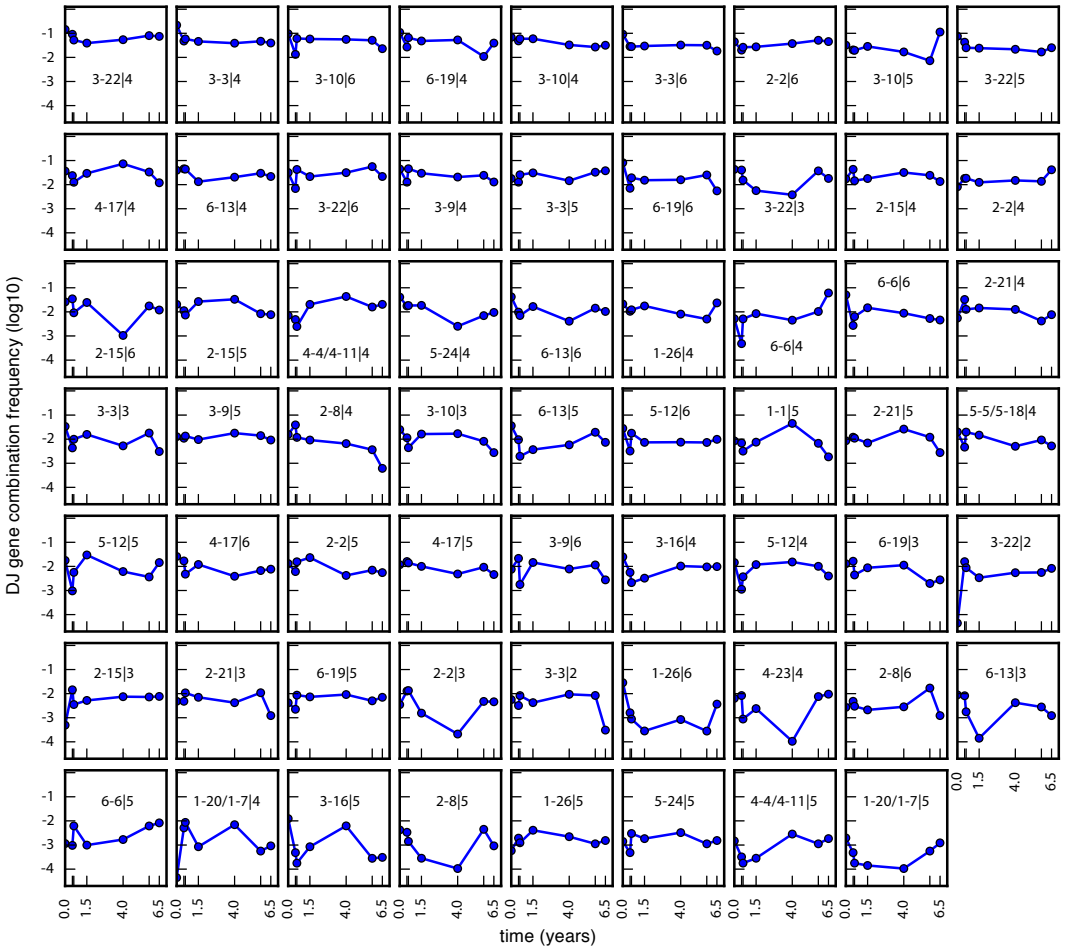


Figure 2.5: **IGH D-J usage frequencies for Donor 1** Plots are sorted by decreasing mean frequency. Only gene identifications that appear in all timepoints are shown.

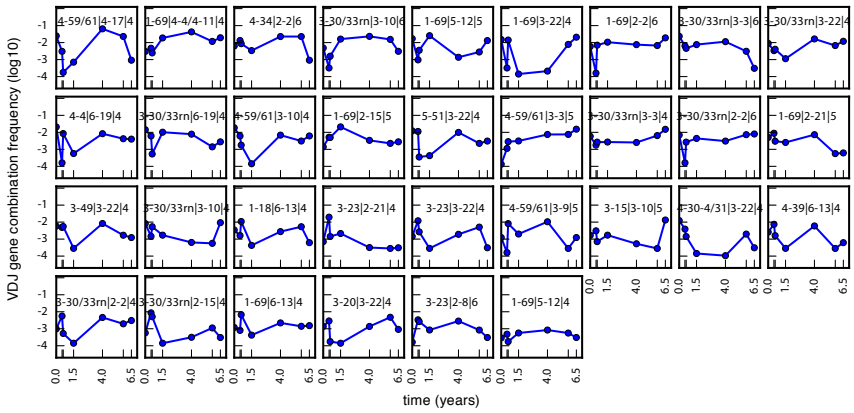


Figure 2.6: **IGH V-D-J usage frequencies for Donor 1** Plots are sorted by decreasing mean frequency. Only gene identifications that appear in all timepoints are shown.

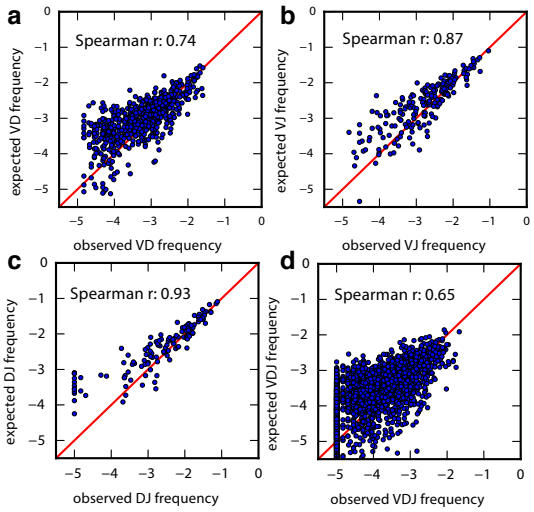


Figure 2.7: **Gene combinations among BM plasma cells do not preferentially associate** Gene combinations are randomly assorted in Donor 1. (a-d) Spearmans rank correlation of expected versus observed IGH V-D (a), V-J (b), D-J (c), and V-D-J (d) gene combination frequencies. Expected (by random association) frequencies are calculated as products of the frequencies of the individual component genes. Diagonal lines in red indicate no difference between the expected and observed frequencies.

identifier of an antibody producing cell. To eliminate errors and ambiguities, we clustered CDR-H3s into clonotypes based on previously established criteria (see Methods). On average, 16% of clonotypes are shared between adjacent timepoints (Figure 2.8a, top). Comparison of the BM plasma cell compartment with memory B cells (mBCs) co-isolated from the same biopsy specimens provided a baseline to gauge stability across the larger framework of the B cell compartment. In mBCs, gene stability was statistically similar to the plasma cell compartment (Figure 2.9). However, no persistent CDR-H3 clonotypes were found among 58,953 mBCs isolated by flow cytometry from the same biopsies across four years in this same donor.

Interestingly, among BM plasma cells, 23 clonotypes persist across all timepoints spanning 6.5 years (Figure 2.8b). We find that 100% of these persistent clonotypes are time stable (Figure 2.8a, bottom, Mann-Kendall test, $h=0$, $p>0.05$) and 78% (18/23) are of the IgA isotype. In addition, characteristics of the complete CDR-H3 population, specifically CDR-H3 lengths (Figure 2.10) and hydropathy index (Figure 2.11a), are unchanged over time. The overall total distribution of CDR-H3 lengths are consistent with previously reported single timepoint values. Also, higher expressing CDR-H3s tend to be neither hydrophobic nor hydrophilic (Figure 2.11b) and we find no significant trends between hydrophobicity and expression level.

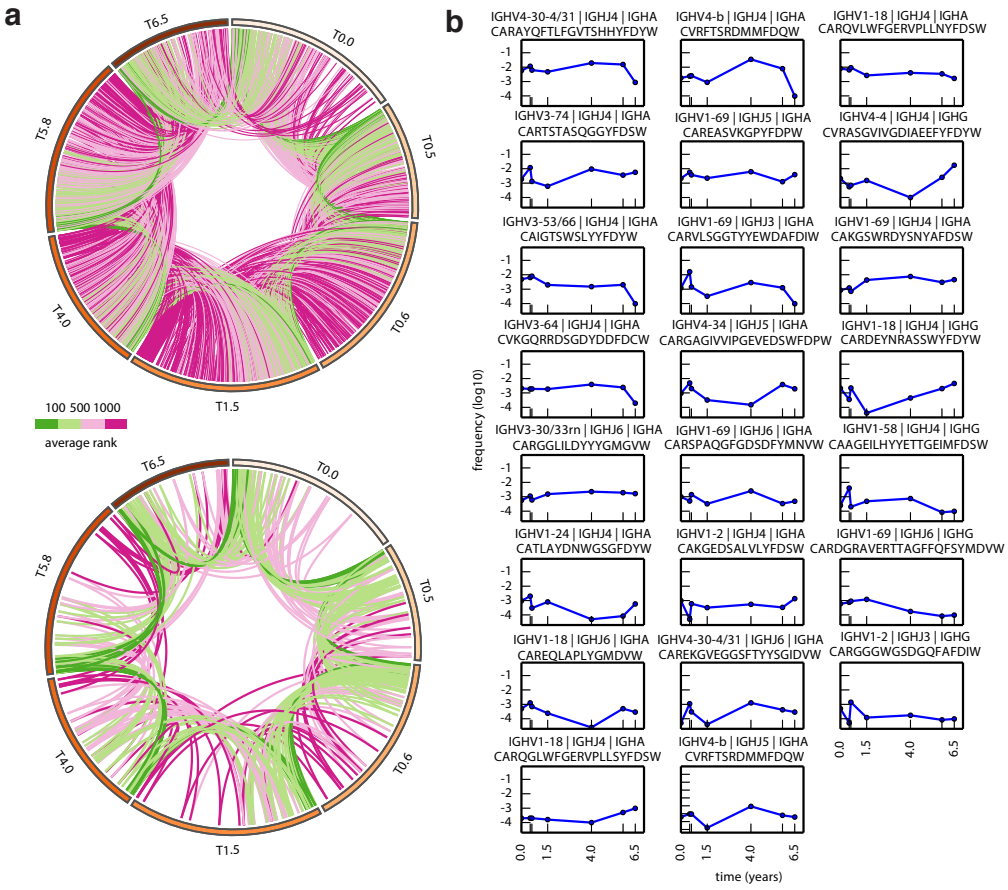


Figure 2.8: **Frequencies of persistent antibody clonotypes among BM plasma cells are temporally stable** (a) Circos plot of shared CDR-H3 antibody clonotypes between adjacent timepoints across 6.5 years for Donor 1 (top). Circos plot of the persistent clonotypes across all timepoints (bottom). Each band in the outermost perimeter represents the clonotypes found in a given timepoint, sorted by decreasing frequency. The inner curved lines indicate the same clonotype shared by two timepoints. Green indicates high frequency; purple, low frequency; with lighter colors indicating intermediate frequency. (b) Gene usage frequency over time of the 23 persistent clonotypes (see Methods) found in all timepoints. Plots are sorted by decreasing mean frequency. Gene names (for IGHV and IGHJ), representative amino acid sequences, and isotype are above each plot.

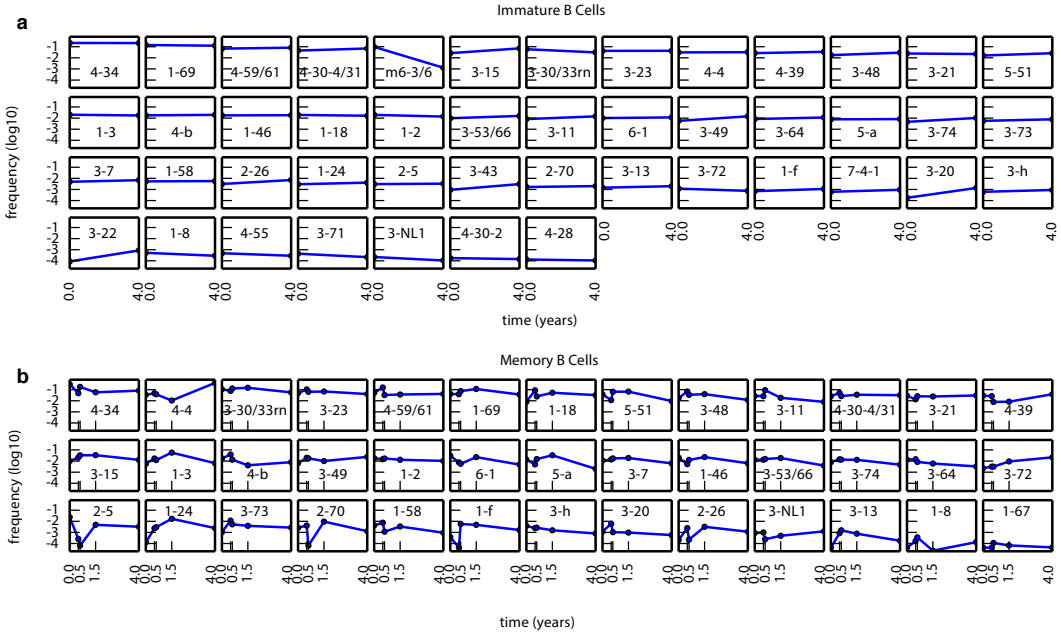


Figure 2.9: **IGHV frequencies across four years in Donor 1 in immature B and memory B cell subsets isolated from bone marrow** Plots are sorted by decreasing mean frequency.

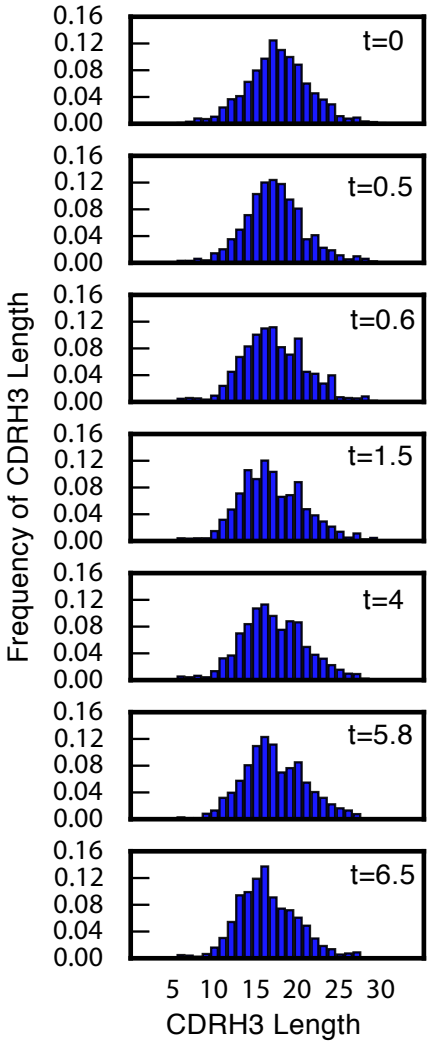


Figure 2.10: **CDR-H3 length distribution for each timepoint from Donor 1**

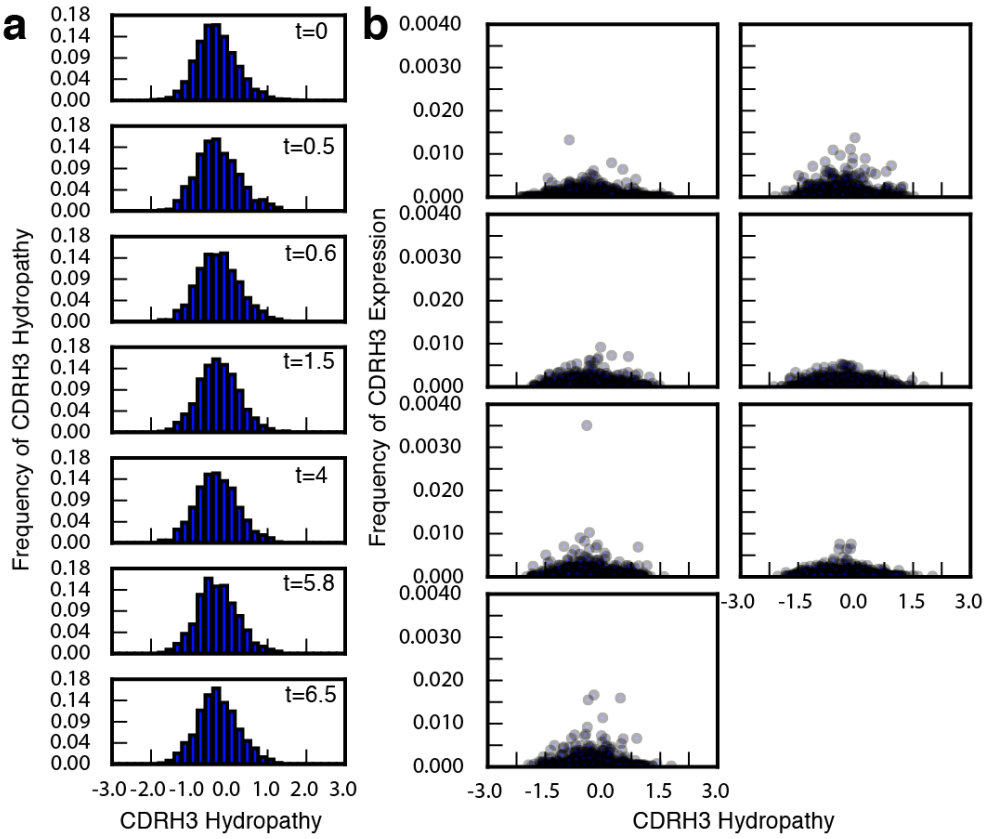


Figure 2.11: **CDR-H3 frequency and hydropathy distribution For Donor 1** (a) CDR-H3 hydropathy distribution for each timepoint. (b) CDR-H3 frequency versus hydropathy scatter plot.

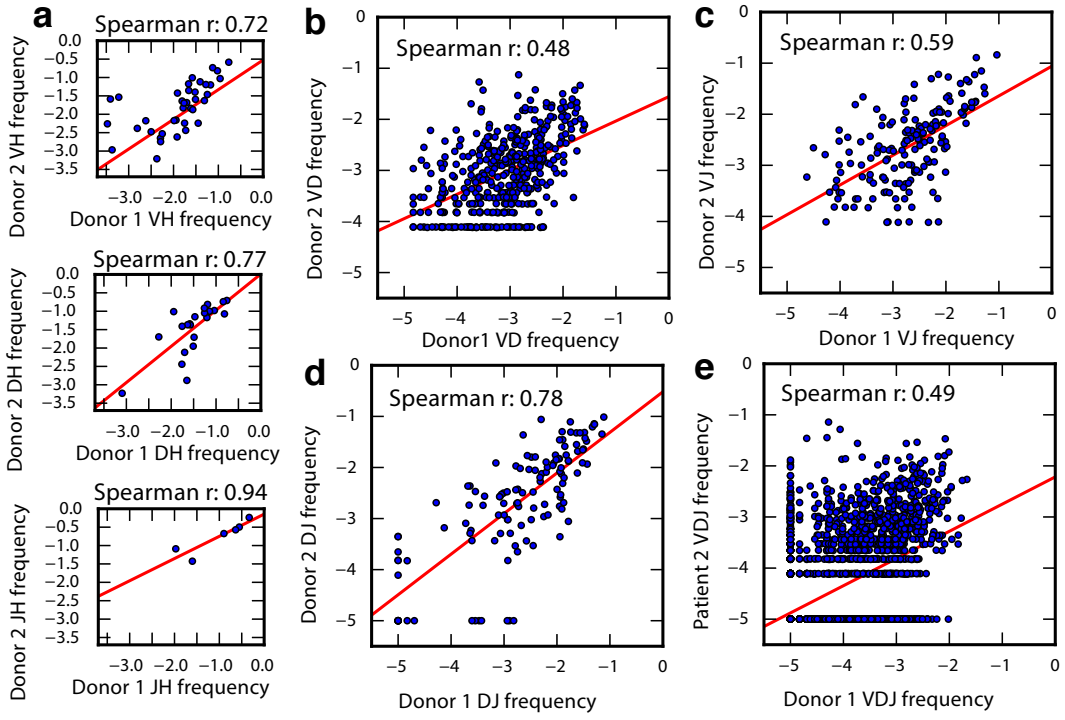


Figure 2.12: **Gene and gene combination use frequencies correlate between Donor 1 and Donor 2** (a) Spearmans rank correlation of individual gene frequencies between the two donors: IGHV (top), IGHD (center), and IGHJ (bottom). (b-e) Spearmans rank correlation of combination gene frequencies between the two donors: V-D (b), V-J (c), D-J (d), and V-D-J (e). (a-e) Red lines indicate least squares regression.

2.3.5 Second donor corroborates observations from first donor

To verify our longitudinal observations of stability and random gene choices from Donor 1, we analyzed a second donor across two years (Figure 2.12). We identified 38 IGHV genes, 22 IGHD genes, and 6 IGHJ genes (5,016 combinations, 6,763 cells, 93,936 reads) (Table 2.1 and Figure 2.13). Donor 1 and Donor 2 show highly correlated IGHV gene usage (r=0.82). Thus, the trends observed in Donor 1 are also observed in Donor 2. Specifically, individual IGHV, IGHD, and IGHJ gene usages are time stable (Figure 2.13), as are the gene combinations (Figure 2.14). Consistent with Donor 1, Donor 2 shows no preferential pairing in gene combinations (Figure 2.15). These results are highly consistent with the trends observed in Donor 1, and together, they indicate that BM plasma cell antibody gene and gene combination usage show surprisingly minimal variation between individuals and across time. Interestingly, minimal variation and a high degree of correlation is maintained when the BM plasma cell repertoires of Donor 1 or Donor 2 are compared with the BM plasma cell repertoire that was obtained from a single donor (age 64) in a separate study [28] (Figures 2.16 and 2.17).

Like Donor 1, no persistent CDR-H3 clonotypes are found among 24,287 mBCs sorted from the same biopsies across 2.3 years in Donor 2. In contrast, persistent CDR-H3 clonotypes (165) are readily detected in the BM plasma cell compartment (Figure 2.18). Importantly, these 165 clonotypes are exclusive to the plasma cell compartment (i.e., absent among mBCs). Lastly, as a measure of the quality and integrity of the B-cell sequence datasets derived from the

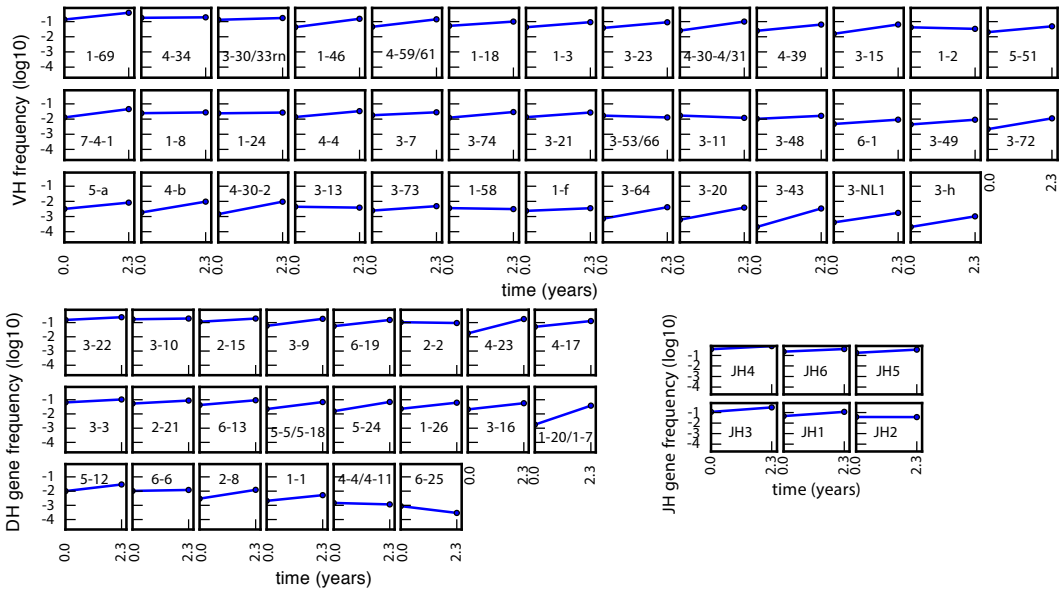


Figure 2.13: **Gene usage frequency over time for Donor 2** For Donor 2: (a-c) IGHV (a), IGHD (b), and IGHJ (c) gene usage frequency over time. Plots are sorted by decreasing mean frequency. Only gene identifications that appear in all timepoints are shown. (d) Mean frequency of IGHV gene use. Error bars are standard deviation.

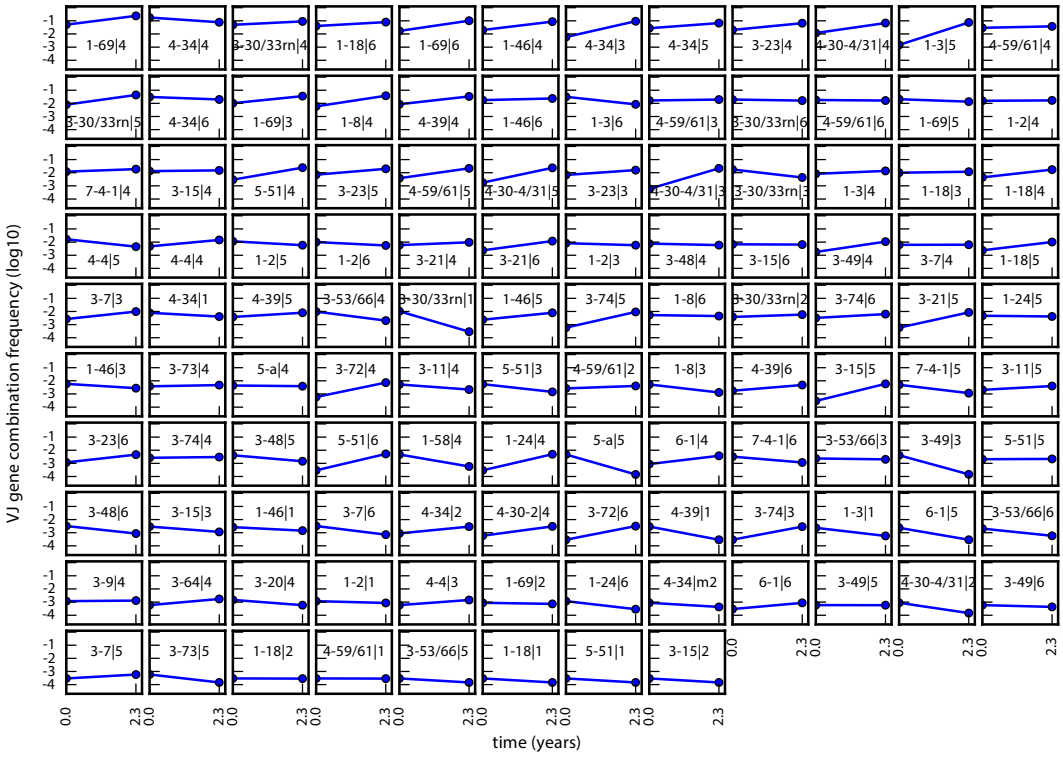


Figure 2.14: **IGH V-J usage frequencies for Donor 2** Plots are sorted by decreasing mean frequency. Only gene identifications that appear in all timepoints are shown.

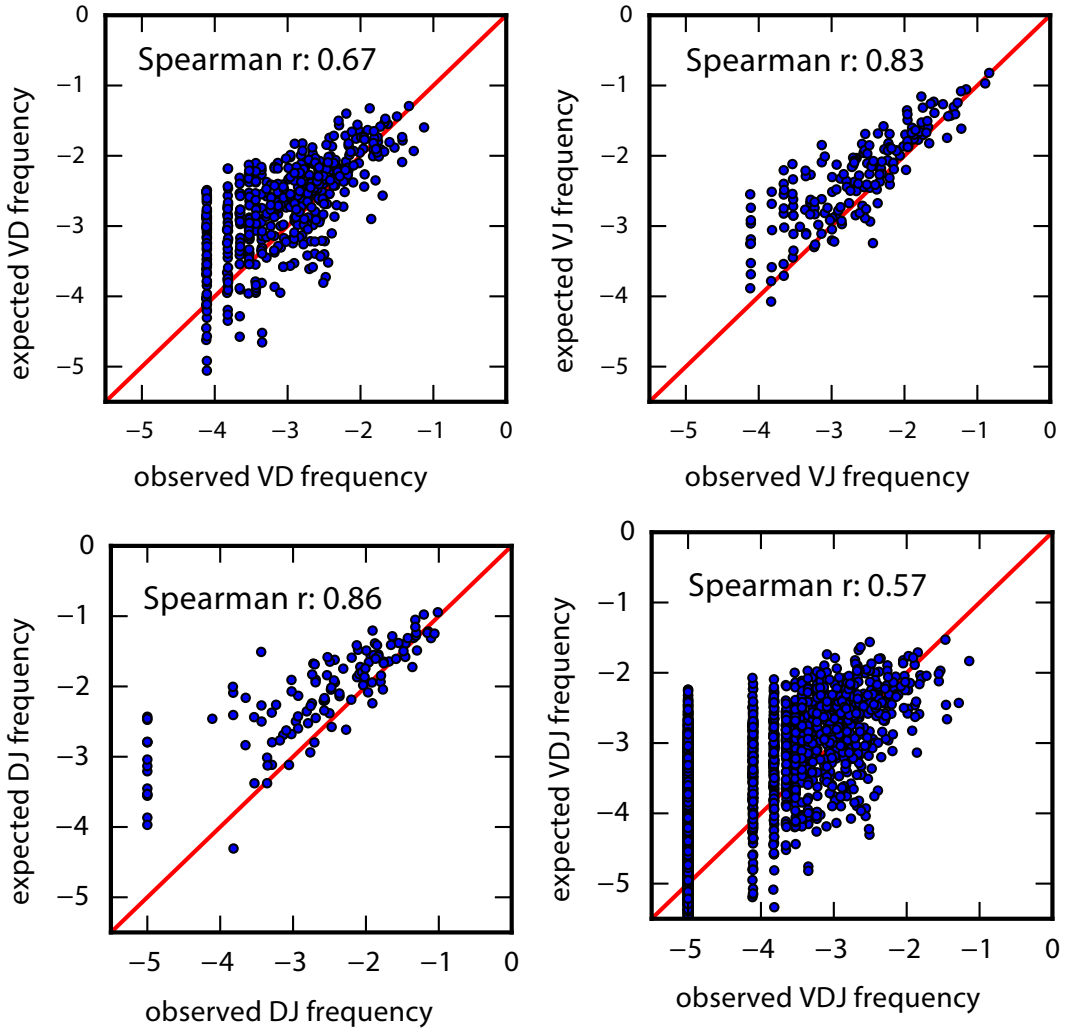


Figure 2.15: **Gene combinations among BM plasma cells are randomly assorted in Donor 2** Gene combinations among BM plasma cells are randomly assorted in Donor 2. (a-d) Spearmans rank correlation of expected versus observed IGH V-D (a), V-J (b), D-J (c), and V-D-J (d) gene combination frequencies. Expected (by random association) frequencies are calculated as products of the frequencies of the individual component genes. Diagonal lines in red indicate no difference between the expected and observed frequencies.

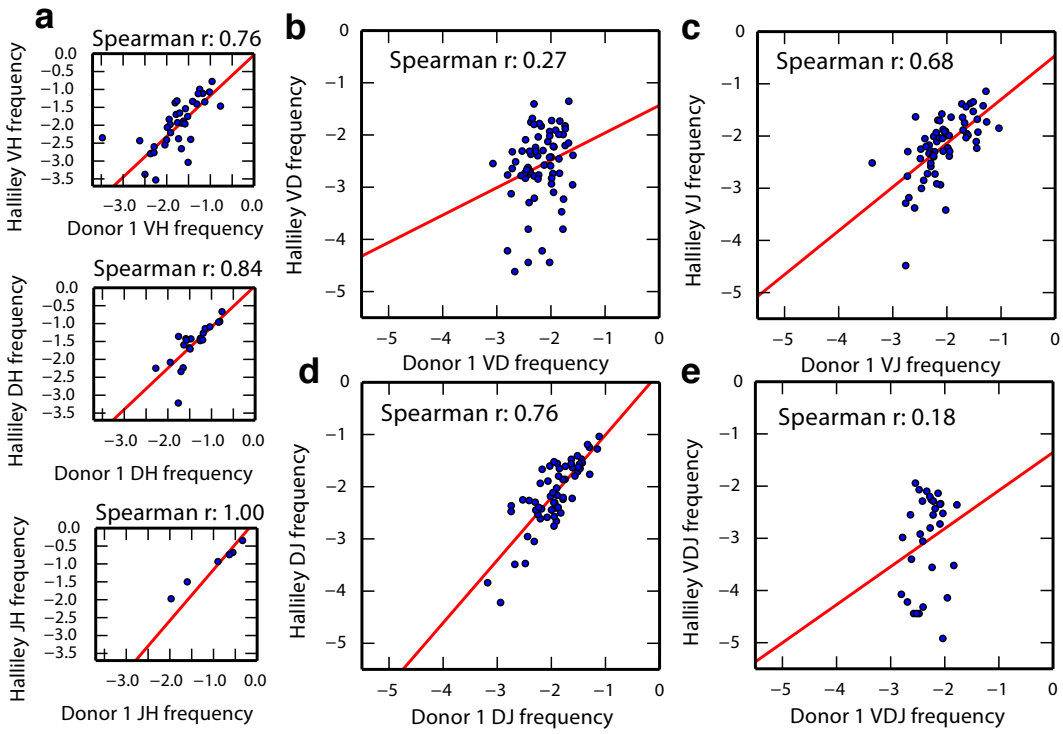


Figure 2.16: **Gene and gene combination use frequencies correlate between Donor 1 and donor from Halliley, 2015** (a) Spearmans rank correlation of individual gene frequencies between the two donors: IGHV (top), IGHD (center), and IGHJ (bottom). (b-e) Spearmans rank correlation of combination gene frequencies between the two donors: V-D (b), V-J (c), D-J (d), and V-D-J (e). (a-e) Red lines indicate least squares regression.

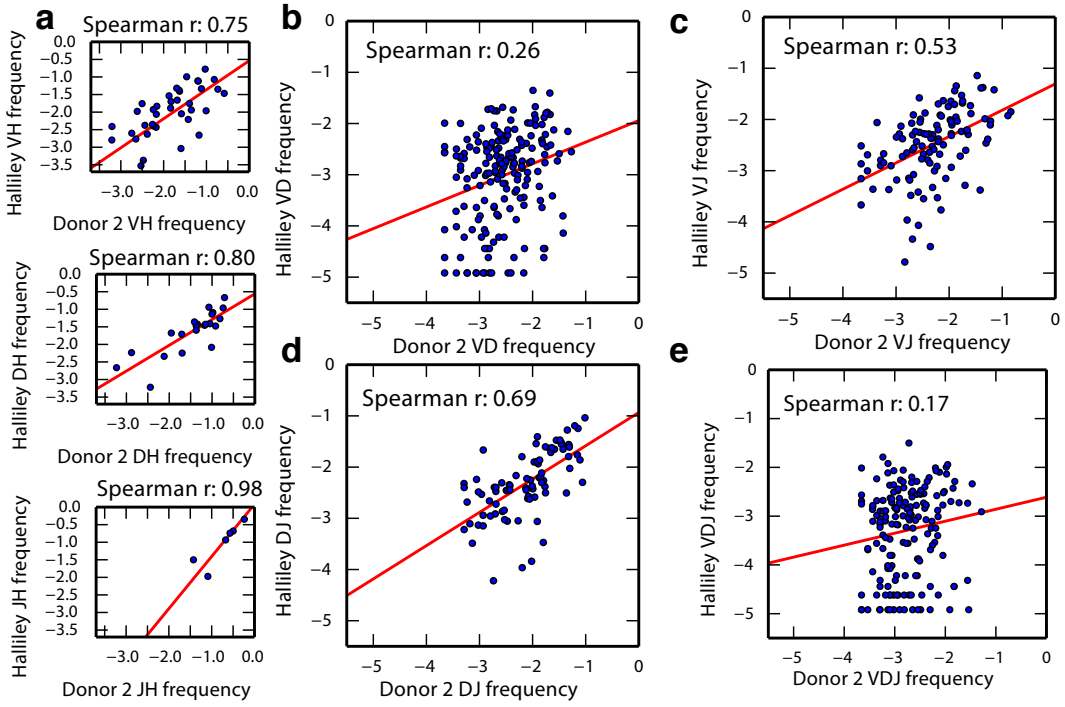


Figure 2.17: **Gene and gene combination use frequencies correlate between Donor 2 and donor from Halliley, 2015** (a) Spearmans rank correlation of individual gene frequencies between the two donors: IGHV (top), IGHD (center), and IGHJ (bottom). (b-e) Spearmans rank correlation of combination gene frequencies between the two donors: V-D (b), V-J (c), D-J (d), and V-D-J (e). (a-e) Red lines indicate least squares regression.

two donors, we observe no inter-donor sequences shared between their mBC compartments, as expected, and only one of the total 188 persistent plasma cell clonotypes is common between the two donors.

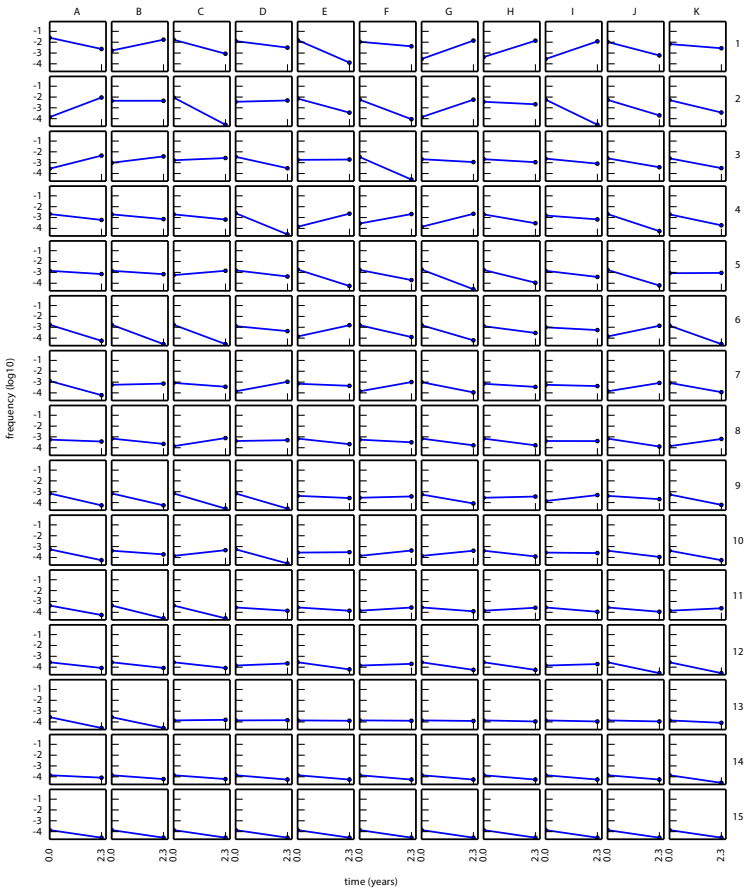


Figure 2.18: **Gene usage frequency over time of the 165 persistent clonotypes found in both timepoints in Donor 2** Plots are sorted by decreasing mean frequency. The gene names (for IGHV and IGHJ), representative amino acid sequences, and isotype information can be found in Table 2.2.

Table 2.2

Position	VH gene	JH Gene	Isotype	Representative CDR-H3
A1	IGHV1-69	IGHJ1	IGHG	CARHPSNSWFRIHFQHW
B1	IGHV1-69	IGHJ1	IGHA	CARGGEQGNYYRTWEYYPYW
C1	IGHV4-34	IGHJ4	IGHM	CARWIRYCSGGDCYPSMYYFDYW
D1	IGHV1-18	IGHJ6	IGHG	CARDRCSGGSCYPGRPQYFYGMDVW
E1	IGHV1-24	IGHJ3	IGHG	CATVAITVDYDSTAYDGLDVW
F1	IGHV1-69	IGHJ4	IGHG	CAKASQNYDSSGYFDCW
G1	IGHV1-3	IGHJ6	IGHA	CARVTATSILGDSGRHHYYAMDVW
H1	IGHV1-46	IGHJ3	IGHA	CARGLRGNLRVLAILPAGAFDMW
I1	IGHV1-46	IGHJ6	IGHA	CARPLSQRGHFYYGMDVW
J1	IGHV4-34	IGHJ4	IGHG	CARGRIVVAPAAMFRRRGSDYFDYW
K1	IGHV1-69	IGHJ6	IGHG	CASDNKIYDYGDGDFQYHNLAVW
A2	IGHV3-15	IGHJ5	IGHG	CVTQATAATAGLAAIITNFDLW
B2	IGHV1-46	IGHJ3	IGHA	CARVIKPGKNDVFEIW
C2	IGHV1-3	IGHJ5	IGHG	CARVVDTPFCRSSNCHNWLDPW
D2	IGHV1-69	IGHJ5	IGHA	CATWGGHCTWYNWCSRVTAFSLDIW
E2	IGHV3-23	IGHJ4	IGHA	CAKAPLDVVTELDYW
F2	IGHV4-34	IGHJ4	IGHG	CARVVNGVAPAAIFHRRGLDYFDYW
G2	IGHV1-69	IGHJ3	IGHA	CARDLRDMSASGGVTFDAFNIW
H2	IGHV1-69	IGHJ4	IGHA	CARWDGHCSFFNWCSGRTVFPLDFW
I2	IGHV1-8	IGHJ4	IGHG	CARGGGSNWRRIHPVDYW
J2	IGHV1-69	IGHJ5	IGHG	CARDMNDYYDPSGYSGALDHW
K2	IGHV4-34	IGHJ4	IGHG	CARARVRNPTGLFRRGYPVFDSW
A3	IGHV4-30-4/31	IGHJ4	IGHA	CAVMYNWNYGFDYW
B3	mIGHV6-3/6	IGHJ4	IGHG	CARYVWYSSYPHSYSGLDYW
C3	IGHV4-30-4/31	IGHJ3	IGHG	CARVGYDGRDYVGKYGFDIW
D3	IGHV4-34	IGHJ4	IGHG	CAGKRRLYSYGLGSYYYFESW
E3	IGHV3-7	IGHJ5	IGHM	CARRGPTFWSGYYESYYDAW
F3	IGHV1-3	IGHJ6	IGHG	CATTNRQIRAARDFYGMDVW
G3	IGHV3-53/66	IGHJ4	IGHG	CARTGQDWYDIHLEHW
H3	IGHV3-30/33rn	IGHJ4	IGHG	CARELYAGSSGYVGYFDSW
I3	IGHV1-69	IGHJ4	IGHA	CATWGGQCAWYNWCNRNTAFSLDFW
J3	IGHV1-69	IGHJ5	IGHG	CALGVKGFMVHGGAKNWFESW
K3	IGHV1-18	IGHJ3	IGHG	CARGTDYGDYIGAFDFW
A4	IGHV1-3	IGHJ6	IGHG	CARVTATSELRDTGRHHYYIMDVW
B4	IGHV3-15	IGHJ6	IGHG	CATGSHPGRKFYYGSVFW
C4	IGHV3-30/33rn	IGHJ6	IGHG	CARDSVHMINSYDYYFGMDVW
D4	IGHV3-30/33rn	IGHJ4	IGHG	CARDCSGYFCFDHW
E4	IGHV4-34	IGHJ3	IGHG	CAACGGSSSCGRAFDIW
F4	IGHV5-51	IGHJ4	IGHG	CARHRGDPFYHGLESRMRFFDYW
G4	IGHV4-34	IGHJ6	IGHG	CARGHDFLSPPGYYYGLDVW
H4	IGHV1-69	IGHJ3	IGHG	CARTRALADGGAFEIW
I4	IGHV3-30/33rn	IGHJ6	IGHG	CAKEESNHVNYYYYYAMDVW
J4	IGHV3-30/33rn	IGHJ5	IGHG	CARYYYDTSGPVLDLW
K4	IGHV4-34	IGHJ4	IGHG	CARLVSVVVPSALFHRRGLEYFDSW
A15	IGHV4-34	IGHJ3	IGHG	CARRVATIARGAFDIW
B5	IGHV3-30/33rn	IGHJ4	IGHG	CARIHISAPGNNFDYW
C5	IGHV1-24	IGHJ6	IGHA	CATGEGDAYNYGLDVW
D5	IGHV3-30/33rn	IGHJ1	IGHG	CARIHIAAHGNNFESW
E5	IGHV4-34	IGHJ4	IGHG	CASFAGFRDKWSHLAYW
F5	IGHV1-18	IGHJ4	IGHG	CARDLKGVSVSATFWGLSDDW
G5	IGHV3-11	IGHJ4	IGHG	CARVHSYGDRGPFDYW
H5	IGHV4-34	IGHJ6	IGHG	CVRGHPYKGLGKLYYHYYYGMDVW
I5	IGHV1-69	IGHJ5	IGHA	CATWGGHCTWYNWCSRVTAFSLDIW
J5	IGHV1-46	IGHJ6	IGHG	CARGDTMVGGIDCMDVW
K5	IGHV1-69	IGHJ6	IGHG	CSRSLRGRWLQSDRDYYYAMDVW
A6	IGHV4-30-4/31	IGHJ4	IGHG	CARVVETATDYW
B6	IGHV3-30/33rn	IGHJ4	IGHA	CARVFESYNLDHW
C6	IGHV4-59/61	IGHJ2	IGHG	CARGRSGDYILYWYLDLW
D6	IGHV1-24	IGHJ5	IGHG	CASIMGHDYGDYVETPNWFDPW
E6	IGHV1-46	IGHJ6	IGHA	CARDPVGATRGGGGMDVW
F6	IGHV1-2	IGHJ3	IGHG	CARGSDRGYAVLGELSAGGAFDIW
G6	IGHV1-2	IGHJ5	IGHM	RATTYCNGVCPDDNWFDPW
H6	IGHV1-2	IGHJ5	IGHG	CARDGRPLQFLKNWFDPW
I6	IGHV4-34	IGHJ6	IGHG	CARMVVKQQLLPRFQVGYYGMDVW
J6	IGHV1-18	IGHJ1	IGHA	CTRDNSNYPEYFQHW
K6	IGHV1-8	IGHJ3	IGHM	CARGSYYDSSGHYHRIAFDIW
A7	IGHV3-30/33rn	IGHJ6	IGHG	CARWAYEGTDVYYYYGMDVW
B7	IGHV1-18	IGHJ5	IGHA	CAKDLWTVTPSFNWFDSW
C7	IGHV1-46	IGHJ4	IGHA	CAREFLGPDYYGSGTKYEYW
D7	IGHV1-69	IGHJ6	IGHA	CARVPYFGSGSYYENYYDMDVW
E7	IGHV1-69	IGHJ6	IGHA	CARLPFFGSGSYYENYYDMDVW
F7	IGHV1-69	IGHJ6	IGHG	CAREGGYCTSPRCYVLEWPRNAGPDYYYNYHMNVW
G7	IGHV1-3	IGHJ5	IGHG	CARSDQWLVGLDPW
H7	IGHV4-34	IGHJ6	IGHG	CARGRFKVVVFGVALEYGLDVW

Continued on next page

Table 2.2 – continued from previous page

Position	VH Gene	JH Gene	Isotype	Representative CDR-H3
I7	IGHV1-69	IGHJ4	IGHA	CATTEDGRVPGYFDYW
J7	IGHV3-30/33rn	IGHJ6	IGHG	CAKDEQMTATYYYYFYGMDVW
K7	IGHV1-69	IGHJ4	IGHA	CVRESRKDGYGRDW
A8	IGHV4-34	IGHJ6	IGHG	CARRYDASGSHYYFYYYHMDVW
B8	IGHV3-30/33rn	IGHJ4	IGHG	CAKDGGIGFTDFDSW
C8	IGHV1-46	IGHJ4	IGHA	CAREGTSRFFQYW
D8	IGHV3-30/33rn	IGHJ1	IGHG	CARIHIRAGGNNFDSW
E8	IGHV4-30-4/31	IGHJ4	IGHG	CARVGPFDTTGYYFDYW
F8	IGHV1-2	IGHJ4	IGHG	CAREAPNLRYYFDFW
G8	IGHV3-30/33rn	IGHJ2	IGHA	CAKDRGISGSYLDWYFDLW
H8	IGHV4-34	IGHJ4	IGHG	CARGVYSGSGSYDYW
I8	IGHV1-69	IGHJ6	IGHA	CAREETEYTTSSLRTTTPYNYGLDIW
J8	IGHV1-46	IGHJ3	IGHA	CARVTKPGKNDVFEIW
K8	IGHV1-18	IGHJ5	IGHA	CARGHIWKELDSW
A9	IGHV1-3	IGHJ6	IGHG	CARDGRGSYGSDFYHSMDAW
B9	IGHV1-69	IGHJ4	IGHA	CARVPTTNILDSGYDYYFDYW
C9	IGHV1-46	IGHJ4	IGHG	CARDISSWHEPRYYFDDW
D9	IGHV1-8	IGHJ5	IGHG	CARVYGVWGVERGLQNQHFDQW
E9	IGHV1-18	IGHJ5	IGHG	CARDTPNYQLLEDFW
F9	IGHV1-18	IGHJ4	IGHA	CTRDTPNYQLLEDYW
G9	IGHV4-39	IGHJ4	IGHG	CTRDSGFYLRMGYW
H9	IGHV3-21	IGHJ4	IGHA	CARGAGGNPVGPTKEPKGGFDYW
I9	IGHV3-30/33rn	IGHJ4	IGHG	CARIHIRAAGNNFDNW
J9	IGHV1-3	IGHJ4	IGHA	CAREGVDMPTVWPIRPSRNYFDSW
K9	IGHV1-69	IGHJ4	IGHA	CARWNGHCSFFNWCSGRTVFPLDFW
A10	IGHV4-34	IGHJ5	IGHG	CARLGVVLPAAMFSRKGGNQFDPW
B10	IGHV4-b	IGHJ4	IGHA	CARGPRTMYNSNYYDYFFDYW
C10	IGHV1-3	IGHJ6	IGHA	CARVTATSIVTDAGRLWYYAMDVW
D10	IGHV1-8	IGHJ4	-	CARGRGAAVVRPETYW
E10	IGHV1-2	IGHJ4	IGHA	CARAWNDVPGGYW
F10	IGHV4-59/61	IGHJ5	IGHG	CARSTLSYCGDSCYPLDSW
G10	IGHV1-18	IGHJ6	IGHG	CVRDIFSTEWTLGYHGMDVW
H10	IGHV4-34	IGHJ5	IGHG	CARLTSVVPAAMFSRMGGDHFDPW
I10	IGHV3-30/33rn	IGHJ3	IGHG	CAREGSGWLAAFDIW
J10	IGHV3-23	IGHJ4	IGHG	CAKKRLVGFLHHFFDSW
K10	IGHV1-69	IGHJ4	IGHM	CARVMEYCSGGSCYEDFDYW
A11	IGHV1-46	IGHJ3	IGHG	CARGVTLYYGESDAGDAFDIW
B11	IGHV1-18	IGHJ5	IGHA	CARDRCITTSCYPWFDPW
C11	IGHV3-53/66	IGHJ6	IGHA	CARAPGLQGGYYYYYGMEVW
D11	IGHV1-18	IGHJ5	IGHA	CARVDFYDLLPGYCKYW
E11	IGHV3-74	IGHJ4	IGHA	CVRSHTGRYDNW
F11	IGHV1-18	IGHJ5	IGHA	CARDLWTVTPSFNWFESW
G11	IGHV1-69	IGHJ5	IGHG	CATWGGHCTWYSWCSRVTAFSLDIW
H11	IGHV4-34	IGHJ6	IGHG	CVRGPREEPAGPSHPRYYFYYSAIDVW
I11	IGHV1-2	IGHJ4	IGHA	CATSLELRVPDDSW
J11	IGHV4-39	IGHJ3	IGHA	CAREDSYKTRNTFDIW
K11	IGHV1-2	IGHJ4	IGHG	CARTLEDYEDYW
A12	IGHV1-69	IGHJ5	IGHG	CARGRDDYKGEVFDHW
B12	IGHV4-34	IGHJ6	IGHG	CARMVIKQQPLPRFQVAYYGMDVW
C12	IGHV4-34	IGHJ4	IGHA	CARGPPGYALDYW
D12	IGHV1-46	IGHJ6	IGHA	CARDFRAILLVRGVLRDYALDVW
E12	IGHV3-23	IGHJ4	IGHG	CAKEDCSSANCYRLDYW
F12	IGHV4-59/61	IGHJ6	IGHA	CARVVTLRVAGSSQYYMDTW
G12	IGHV1-3	IGHJ6	IGHG	CARVTATSRVTDAGRLWFYAMDVW
H12	IGHV4-59/61	IGHJ4	-	CAVNYDSSGYTRGFDSW
I12	IGHV1-69	IGHJ3	IGHG	CARDGGYCSGRACHAYAFDMW
J12	IGHV1-69	IGHJ6	IGHG	CARDIAVSETDYYYFALDVW
K12	IGHV3-30/33rn	IGHJ4	IGHA	CASELTRVAAAGKGNDYW
A13	IGHV4-59/61	IGHJ3	IGHG	CARPIWEPRDAFDIW
B13	IGHV1-3	IGHJ1	IGHA	CARRPYCSGGSCYTGEYFQHW
C13	IGHV1-8	IGHJ5	IGHA	CARGNKPDHTASSLSKNWFDPW
D13	IGHV1-18	IGHJ6	IGHA	CARDDRYSSAWYLGSYYGMDVW
E13	IGHV4-39	IGHJ5	IGHG	CARHYDFVWGTYRDQARNWFDPW
F13	IGHV5-51	IGHJ3	IGHA	CARPEAISGFYAFDVW
G13	IGHV1-69	IGHJ5	IGHA	CARWDGHCSFFNWCSGRTVFPLDFW
H13	IGHV1-69	IGHJ4	IGHA	CASAGGDDIFAVVTYYW
I13	IGHV3-11	IGHJ4	IGHM	CARGLRGYSYGLSDYW
J13	IGHV4-4	IGHJ4	IGHM	RASRRVGATFYW
K13	IGHV1-18	IGHJ4	IGHA	CARVQSNSIFGVFIPYHLDSW
A14	IGHV4-34	IGHJ5	IGHA	CARWIRYCSGGDCYPSMYYFDSW
B14	IGHV1-2	IGHJ6	IGHA	CFRETQRGYGMDVW
C14	IGHV3-15	IGHJ6	-	CATGSHPGRKVLHGSVVW
D14	IGHV1-69	IGHJ4	IGHA	CARESGDGYNPKRAHVFDYW
E14	IGHV1-69	IGHJ3	IGHA	CASHQPKNYYDSSSYRAFDIW

Continued on next page

Table 2.2 – continued from previous page

Position	VH Gene	JH Gene	Isotype	Representative CDR-H3
F14	IGHV1-3	IGHJ5	IGHG	CAREPVPHQLLYWFDPW
G14	IGHV4-34	IGHJ4	IGHG	CARGRIVVASAALFRRRGSDYFDYW
H14	IGHV4-34	IGHJ4	IGHA	CARLVSVVQPAALFHRRGLDYIDFW
I14	IGHV1-18	IGHJ1	IGHG	CARGHIWKELDSW
J14	IGHV3-48	IGHJ5	IGHM	CALSRDGYSHKW
K14	IGHV4-59/61	IGHJ6	IGHM	CARRSGGSHYYMDVW
A15	IGHV4-34	IGHJ6	IGHA	CVRGHPYKGFGEKYYLYYYGMDVW
B15	IGHV4-34	IGHJ4	IGHG	CARGQTALKPVVFGVVITRPTNNYFDYW
C15	IGHV3-21	IGHJ4	IGHA	CARDDGDSVAEEYW
D15	IGHV4-34	IGHJ5	IGHG	CARLGVVVPVAMFSRKEGNHFDPW
E15	IGHV1-2	IGHJ6	IGHA	CARDFLPPGQVATIPLWHGMDVW
F15	IGHV4-34	IGHJ6	IGHM	CARGHEDYSNYYYYGMDVW
G15	IGHV1-8	IGHJ6	-	CARVGGPYSIHYMDVW
H15	IGHV1-69	IGHJ6	IGHG	CARDGRGQRPTRHIIIINTDWYLW
I15	IGHV1-69	IGHJ4	IGHG	CARSPVAGAYFFDYW
J15	IGHV1-69	IGHJ1	IGHG	CARGGNRGVIIGPGNTYPYW
K15	IGHV4-30-4/31	IGHJ4	IGHG	CARGAYFYGSGLDYW

Table 2.2: **Gene names, representative amino acid sequences, and isotypes for persistent CDR-H3s in Donor 2** The IGHV and IGHJ gene names, representative amino acid sequences, and isotype information for Figure 2.18.

2.4 Discussion

High-throughput sequencing has enabled unprecedented ability to explore the details of the human B cell repertoire [23, 41]. Whereas previous studies have been able to describe some aspects of the B cell repertoire at a single point in time, our study harnesses the power of high-throughput sequencing and longitudinal biopsies of bone marrow (BM) to elucidate the temporal dynamics of BM plasma cells over 6.5 years. Importantly, our data provide molecular resolution of antibody identity in the form of CDR-H3 clonotypes, which is not possible with classic techniques like enzyme-linked immunosorbent assay (ELISA).

In this study, we show that the human plasma cell compartment is naturally polarized in both IGH gene choice and gene combination and that the polarization is maintained over time. Although our donors were originally diagnosed with and treated for neuroblastoma, they had been asymptomatic and disease-free for several years, and it was during this span when their BM biopsies were acquired; moreover, it is noteworthy that their IGH polarization is statistically similar to a distinct high-throughput sequencing dataset obtained independently by another research group [28] using a single donor at a single point in time. We also found that the bias is not primarily a result of gene linkage, suggesting there are additional genomic or extrinsic factors that contribute to polarization. Specifically, high-throughput sequencing of identical twin pairs [24, 53] has revealed clear trends for genetic, or heritable, determinants of IGH gene segment use. Nonetheless, the CDR-H3 region maintains

hypervariability and fingerprints inter-individual variation that distinguishes twin pairs. In addition, the long arms race between the human immune system and the antigens it has confronted throughout evolutionary history may have established a preferential gene choice long ago, and thus there may exist common antibody-mediated solutions to protective immunity. Higher expressing genes are likely broad-spectrum antibodies that have been useful in fighting particular classes of disease and continue to do so today. For example, IGHV1-69 is repeatedly implicated in deep sequencing studies of anti-viral antibody repertoires (e.g., influenza and HIV-1), and the inherently autoreactive IGHV4-34 element is associated with a range of autoimmune disorders (e.g., cold agglutinin disease and systemic lupus erythematosus). Indeed, convergent, or public, responses using these IGHV gene segments coupled within homologous CDR-H3 clonotypes continue to be discovered [29, 56, 33].

Immunological memory is a well-established concept, and memory B cells (mBCs) and BM plasma cells are thought to be key contributors, in part, through their putative cellular longevity and hypothesized capacity for self-renewal. How intrinsic longevity might be established and maintained remains an outstanding question. It has been proposed that mBCs generate plasma cells for the lifetime of the human host. It is further hypothesized that mBCs are endowed with a stem cell-like capacity for self-renewal and could be the basis for the continual production of plasma cells [19]. Evidence in support of this hypothesis includes the demonstration that polyclonal activation of mBCs results in their differentiation into plasma cells *in vitro* [7]. Since

class-switched mBCs coexist with plasma cells in human bone marrow [50], we sequenced both compartments to test the hypothesis that BM mBCs may be a renewable source of plasma cells and, indirectly, the source of long-term antibody production in humans. Steady usage of IGH gene and gene combinations observed in both donors throughout our experiment suggests that there are large resident pools of plasma cells of the same identity, from which we can sample continuously with no loss of relative expression levels. Most importantly, we observe years-long temporal persistence of 188 unique, highly diverse CDR-H3 clonotypes exclusively within the plasma cell compartment, whereas CDR-H3 temporal persistence was devoid in the mBC compartment. This provides a crucial point of comparison between these two B-cell subsets pivotal to immunological memory. The data imply that the molecular sequence stability in the plasma cell compartment is due to persistence of the cellular clonotype. It is not simply a reflection of the naïve B cell repertoire nor mBC repertoire in general (i.e., heritable influences of IGH gene use, nor replenishment from the mBC compartment).

Whereas ample data have established the persistence of antigen-specific serum immunoglobulin titers, which have half-lives of decades or longer [2], there is to date no insight as to whether the molecular composition of these antibody titers is a homogenous pool of immunoglobulin maintained by a handful of long-lived plasma cell cellular clonotypes or is rather a continual flux and turnover of transitory plasma cell clonotypes. Although our results are unable to verify the lifespan of any one particular plasma cell, we can conclude that

clonal members of the CDR-H3 clonotype, which defines identity and binding specificity at the molecular level, do persist for at least 6.5 years. Our results suggest that clonotype persistence contributes to the mechanism underlying long-term immunological memory.

In conclusion, we use high-throughput, next-generation sequencing to definitively identify long-term persistent BM plasma cell clonotypes, which has implications in clinical intervention studies, vaccines, and immunotherapy. Future next-generation sequencing studies can provide an even more detailed picture of the B cell immune repertoire including advances in VH:VL native-pair sequencing (paired BCR-Seq [15, 16]), analysis of correlations between BM plasma cell repertoires and serum immunoglobulin species (Ig-Seq [62, 40, 41]), and examination of the connectivity of B cells at various developmental stages (e.g., clonal relationships between circulating memory B cells and sessile BM plasma cells). Our study provides a foundation upon which these further studies can be built.

2.5 Methods

2.5.1 Bone marrow specimens

Serially acquired human bone marrow specimens were collected from two donors by aspiration from the ileac crest, and mononuclear cells were enriched by Ficoll hypaque centrifugation. The two adolescent-teenage donors (10-17 years of age) were originally diagnosed with neuroblastoma but had been asymptomatic and disease-free for many years according to routine bone

marrow histology before the first timepoints in our study. A complete description of the donors past medical history and ages at the time of the multiple time point collections is included in Table 2.1. Aspirates were withdrawn from four sites and combined (total of 810 mL from 4 sites, 22.5 mL per site) drawn from the following: anterior right iliac crest, anterior left iliac crest, posterior right iliac crest, and posterior left iliac crest. The same attending physicians performed these procedures and usually biopsied through the same surgical site each time. De-identified specimens were shipped overnight on dry ice to the University of Texas at Austin.

2.5.2 Flow cytometry and isolation of plasma cells

BM samples were quick-thawed in a 37°C H_2O bath and slowly diluted into RPMI-1640 complete medium containing DNaseI (Sigma D 4513; 20 U/mL), pelleted, washed and re-suspended in 2 mL FACS buffer (Dulbeccos PBS + 0.5% BSA Fraction V). Cell viability was determined using Trypan Blue exclusion and on average was approximately 90% per specimen. After a one-hour recovery at room temperature, BM cells were stained for 30 minutes at room temperature using empirically-determined optimal titrations of monoclonal antibodies: CD38-FITC (HIT2), CD138-PE (B-B4), CD27-APC (M-T271), and CD19-v450 (HIB19). CD19+/-CD38++CD138+ cells in human BM were collected as plasma cells. Plasma cells were observed to be heterogeneous for expression of the CD19 B-lineage marker; therefore, CD19-gating was avoided. CD38++CD138+ plasma cells were additionally gated by

light scatter properties (FSC v. SSC) to exclude debris, apoptotic cells, and remnant granulocytes. In a subset of bone marrow specimens, memory B cells (mBC) were also collected as CD19+CD27+CD38-CD138-. Donor 1 included mBCs at 0, 0.5, 0.6, 1.5, and 4.0 years; Donor 2 included mBCs at 0 and 2.3 years. All cell sorts were performed on a FACSAria (BD Biosciences). Cells were sorted directly into TRI Reagent for RNA preservation.

2.5.3 RT-PCR and high-throughput sequencing of IGH genes

Total RNA was isolated using the RNeasy Micro Kit (QIAGEN). Approximately 100 nanograms of total RNA was then used to prepare oligo-dT primed cDNA using the SuperScript III First-Strand Synthesis System (Thermo-Fisher Scientific) according to the manufacturers protocol. Approximately 25%-50% (5-10 μl) of cDNA was then used as template for polymerase chain reaction (PCR) amplification of variable genes (recombined VHDJH region, which encodes the V region) of IGH isotypes IgM, IgG, and IgA. PCR primers have been published [31]. FastStart High Fidelity PCR System (Roche) was used for amplification combined the following thermocycler conditions: 92°C denaturation for 3 min; 92°C 1 min, 50°C 1 min, 72°C 1 min for 4 cycles; 92°C 1 min, 55°C 1 min, 72°C 1 min for 4 cycles; 92°C 1 min, 63°C 1 min, 72°C 1 min for 20 cycles; and a final extension of 72°C for 7 minutes. Samples were then submitted to the University of Texas Genome Sequencing and Analysis Facility for library construction using the NEBNext Quick DNA Library Kit for 454 (New England Biolabs) and next-generation

sequencing (NGS) was accomplished using the Roche 454 GS FLX technology using titanium long-read chemistry. Read counts per sample are listed in Table 2.1.

2.5.4 Data processing and visualization

IGHV, IGHD, IGHJ, and CDR-H3 regions for each read was quality filtered, processed and annotated using the VDJFasta utility [25]. Reference IGHV, IGHD, and IGHJ genes from the international ImMunoGeneTics (IMGT) database [1] were used. Mann-Kendall Tests were performed in Matlab, against the null hypothesis of no trend ($\alpha = 0.05$). Spearman r non-parametric correlation analysis was performed in Python using the scipy library. CDR-H3 sequences were clustered to form antibody clonotypes, as established previously [67, 12], using full-length VHDJH gene nucleotide sequences. VHDJH genes were grouped into clonotypes based on single-linkage hierarchical clustering, and cluster membership required 85% identity across the CDR-H3 amino sequence (as measured by Levenshtein edit distance).

Circular visualization plots were created with Circos software v0.67-7 [37] where genes were sorted by expression within each timepoint and connected to adjacent timepoints via colored lines showing their expression levels. All other data visualization was performed using Python and matplotlib.

2.5.5 Data availability

All sequence data have been deposited to NCBI SRA under BioProject number PRJNA310043.

2.5.6 Ethics approval

All procedures were performed with parental consent at the Memorial Sloan-Kettering Cancer Center under a protocol approved by the MSKCC Institutional Review Board. The protocol is registered at ClinicalTrials.gov (NCT00588068).

2.5.7 Funding

This work was supported by NSF Graduate Research Fellowship DGE-1110007 (to GCW), HDTRA1-12-C-0105 from DTRA (GG, GCI, EMM), WHO GPEI (GCI), and grants from the NIH, NSF, and Welch Foundation (F-1515) to EMM.

Chapter 3

Computational Simulation of Error in Immune Repertoire Sequencing

3.1 Abstract

While the advent of high-throughput sequencing makes accessible previously intractable questions regarding the immune repertoire, the nature and consequences of errors in the multistep experimental and computational manipulations are not well understood. It remains an open question whether the true biological antibody repertoire of the tissue sampled is accurately captured in the empirically observed repertoire. There lacks a tractable experimental approach for understanding the extent of error generated by the high-throughput sequencing pipeline. Here, we present a computational simulation that is capable of addressing these problems. We demonstrate that there are multiple sources of error, including cell sampling, nucleic acid amplification, and high-throughput sequencing. Under reasonable assumptions, we find that 10-18% of cell sampling is required to capture a majority of the antibody diversity. PCR and sequencing contribute most to the overall error. We quantify the extent of error correction that can be achieved by sequence clustering and demonstrate that clustering is critical for a more accurate reconstruction of the true underlying repertoire. Surprisingly, different true biological distributions,

when processed through our simulation of HTS immune repertoire workflow, result in the same empirical distribution due to the cumulative workflow errors. Based on these results, we provide concrete recommendations to others in the field wanting to perform immune repertoire studies.

3.2 Introduction

Antibody diversity allows the immune system to protect the host from an enormous number of antigens, many of which are pathogenic. This diversity is only possible because of the immune systems ability to rearrange on the genetic level at the post germ-stage of development, a process known as somatic recombination. This process is highly active in B cells that rearrange three genes (V, D and J) in order to produce antibodies. Understanding antibody diversity may enable us to better understand how our immune system responds to external stimuli and the full extent to which it can maintain our health. However, antibody diversity may theoretically be as high as 10^{11} [20], making the full scope and nature of the antibody repertoire difficult to observe. The emergence of high-throughput sequencing (HTS) has led to significant advances in our understanding of the immune repertoire and its diversity. In zebrafish, it was shown that up to 86% of all possible VDJ combinations were used [60]. In humans, complete repertoires have not been studied, but numerous studies have still made significant advances toward that end. For example, preferential V and J gene usage in response to an influenza vaccine has been observed [33]. In addition, VH and D gene use in antibodies have been shown

to be strongly biased by genetics and resistant to chronic lymphocyte depletion [24]. In another study, long lived plasma cells in the bone marrow were shown to have different VH gene usage and RNA transcriptomes in contrast to their non-long lived counterparts [28]. Recently, gene and gene combination usage in bone marrow plasma cells were shown to maintain their bias over time [64]. HTS experiments ultimately result in frequency distributions of immune sequence identities. These distributions are highly polarized, where a small number of sequences account for a disproportionate amount of the total repertoire. The shape of this distribution has been described as a power law [60, 17], more specifically, Zipf's law [49, 27] or in some cases as a Poisson distribution [5, 52]. In order to describe and compare these frequency distributions, two key metrics of diversity are used: species richness and species evenness. Species richness is the total number of unique species in the system. Species evenness measures the degree of polarization in a distribution. Researchers tend to emphasize one of these metrics over the other resulting in metrics that attempt to blend these two properties together into a single number. Specifically, the Shannon Index, Simpson Index, and Berger-Parker Index are common metrics of diversity. A recent proposal attempts to unify many of these most commonly used diversity indices along a single spectrum with species richness on one extreme and species evenness on the other [27]. This can be formalized as:

$$ {}_{\alpha}D(f) = (\sum_{i=1}^{n} f_i^{\alpha})^{\frac{1}{1-\alpha}} \tag{3.1} $$

Where f is the clonal frequency distribution with f_i being the frequency of each clone and n the total number of clones.

In this equation, when $\alpha = 0$, ${}_{\alpha}D(f)$ is species richness, when $\alpha = 1$, ${}_{\alpha}D(f)$ is Shannon entropy index, when $\alpha = 2$, ${}_{\alpha}D(f)$ is the Simpson entropy index, and when $\alpha \to \infty$, ${}_{\alpha}D(f)$ is Berger-Parker.

Observations about the HTS-generated empirical distributions and diversity measures have been used to infer properties of the underlying biological repertoires. However, it is still unclear whether these empirical distributions are representative of true biological distributions. This is because antibody repertoire sequencing poses unique problems to HTS. In contrast to genomic sequencing experiments, HTS antibody experiments cannot rely on a reference sequence, and high fold sequence coverage is difficult to obtain. In addition, short reads are insufficient to accurately identify antibodies because they are the product of multiple gene recombination, junctional diversity, and short hypervariable regions separated by long highly similar sequences. These traits necessitate long accurate reads that push the boundaries of HTS abilities.

Furthermore, because of their complex biology, antibodies exacerbate existing challenges of HTS experiments. First, the immense size and diversity of the immune repertoire result in dramatic undersampling of the cellular compartment. Next, polymerase error from both reverse transcriptase and DNA polymerase generate sequence mutations. Finally, error in the high-throughput sequencer is much higher than by traditional Sanger sequencing.

Here, we present a computational simulation that aims to understand how each of these dependent processes contributes to the overall error of the antibody repertoire HTS experiment. In addition, we asked: given an empirical distribution, could we computationally deduce the true biological distribution from which it originated? To that end, we explore the changes in a theoretical input distribution when subjected to simulations of previously reported antibody repertoire HTS experimental procedures. We offer recommendations on how best to mitigate the errors inherent in the process and suggest the current limits of interpretation of antibody repertoire HTS experiments.

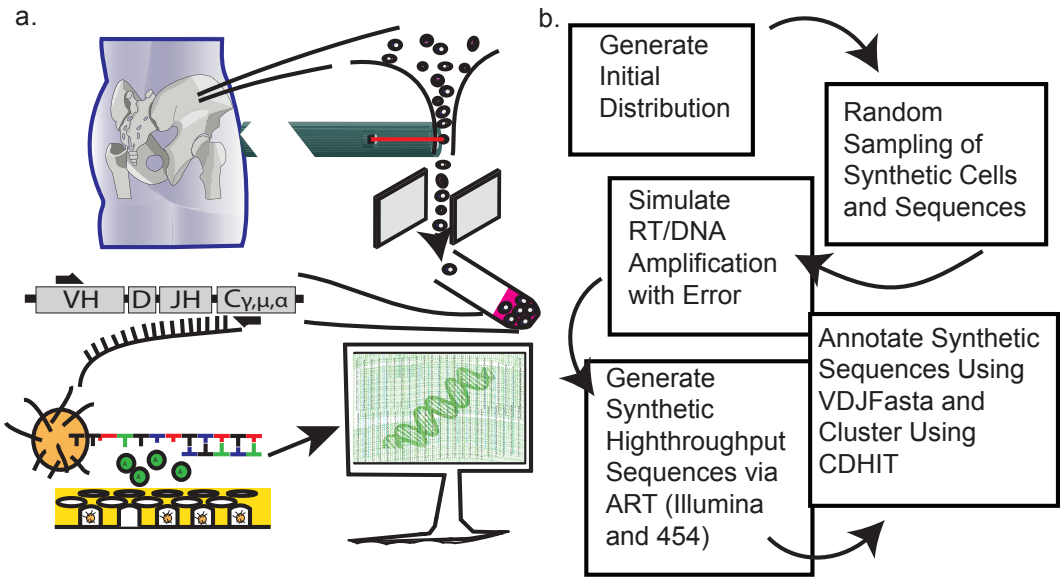


Figure 3.1: **Experimental and model methodology overview** (a) Experimental workflow begins with BM plasma cell sampling followed by FACS sorting, PCR amplification, 454 sequencing and subsequent analysis. (b) Computational modules that simulate experimental steps. (c) Empirical distribution of clonotype counts obtained from workflow depicted in (a)

3.3 Results

3.3.1 Computational simulation of experimental workflow

Our computational simulation aims to provide a theoretical framework for a typical experimental workflow for an antibody repertoire HTS experiment, as depicted in Figure 3.1a. In a typical profiling experiment, samples are first collected from either peripheral blood or bone marrow. Subsequently, B cells are isolated by fluorescently activated cell sorting (FACS). Next, RNA is purified and made into cDNA using reverse transcriptase followed by amplification by polymerase chain reaction (PCR). Then, high-throughput sequencing is performed using pyrosequencing or Illumina sequencing technology. Finally, the data from the sequencer is analyzed and visualized computationally.

In order to understand the errors generated at each of these experimental steps, we computationally simulated each step in Figure 3.1b, using separable computational modules that modeled the errors arising from each of the major experimental manipulations. Because the true biological distribution of the immune repertoire is as yet still unknown, we assumed different theoretical initial distributions of the immune cell identities, i.e. the frequency of unique antibody sequences. We then simulated the cell sampling, amplification, high-throughput sequencing, and data analysis pipeline under different parameters in order to observe how different points of intervention changed the distribution.

3.3.1.1 Theoretical biological distributions

First, we considered possible forms that the true underlying repertoire might take. We analyzed three theoretical distributions that are widely prevalent in nature, testing each in turn as the starting biological distributions in our simulation: exponential, normal, and uniform. In all three distribution types, we used 10,000 total cells and 3,000 unique sequences. The three distributions varied in their levels of polarization and parameterization. For the most polarized distribution, the exponential distribution, the values 0.1, 0.3, and 0.5 were used for the rate parameter, λ. For the normal distribution, varying of the mean parameter did not affect the degree of polarization and the simulation results; a mean value of 10 was chosen for all normal distributions. Also, variance values of 0.2, 0.5, 1, and 5 were used to vary the polarization of the normal distribution. Finally, the uniform distribution relies solely on the number of total cells and unique sequences. The distribution is inherently unpolarized and no additional parameters were used.

3.3.1.2 Mixed Diversity Index

In order to compare across different distributions, we used a mixed diversity index (MDI), which avoids emphasis of any single diversity metric. The MDI is calculated as the average of ${}_{\alpha}D(f)$ [27] for α values from 0-10, since ${}_{\alpha}D(f)$ does not change dramatically for α values larger than 10 [27]. Species Richness (SR), the number of unique individuals, was used in cases where ${}_{\alpha}D(f)$ was divergent for small values of alpha.

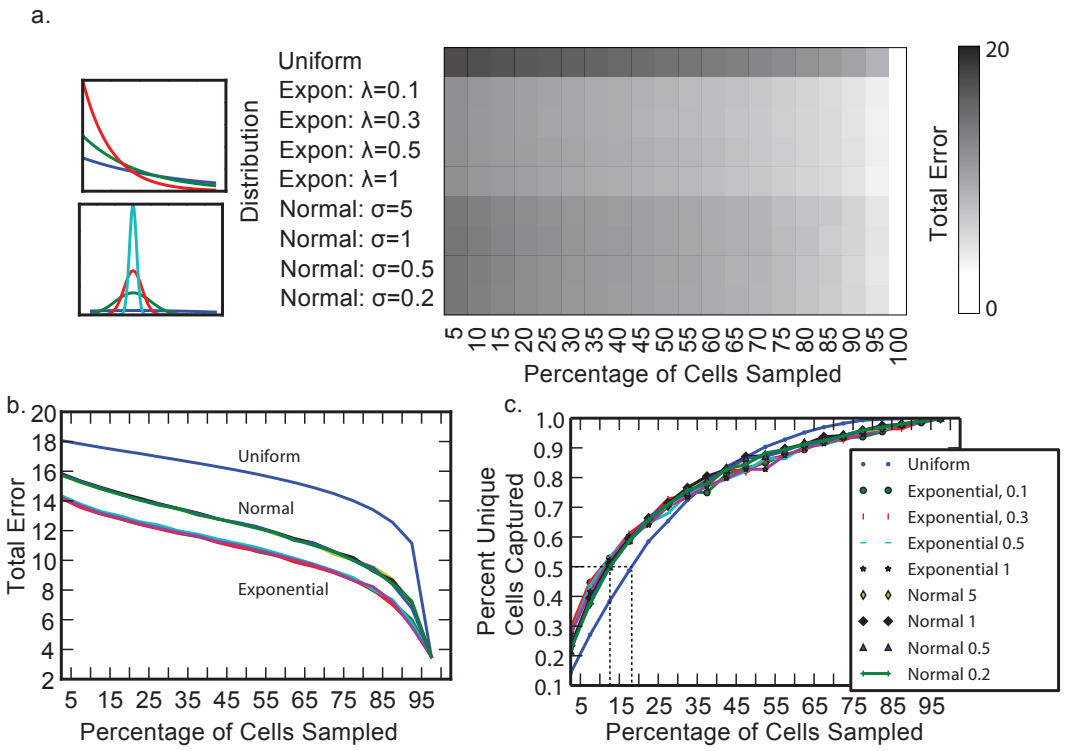


Figure 3.2: **Simulated errors resulting from cell sampling with different initial distributions** (a) Exponential and normal initial distributions used as inputs into the model (left). Total error, defined as the difference between post and pre-sampling MDI (Mixed Diversity Index), resulting from a range of sampling percentages (0-100%). (b) Total error resulting from a range of sampling percentages of various initial distributions. Uniform: dark blue, Normal: green, Exponential: red (c) Percent of diversity captured, measured using percent of unique clonotype IDs represented, over a range of sampling percentages (0-100%). Percent sampling needed for half maximal of species richness are represented by dotted lines.

3.3.2 Cell sampling bias distorts the initial distribution

In order to test the effect of cell sampling on the final distribution, we sampled from the true theoretical biological distributions described above. We found, unsurprisingly, that with increasing percentage of the 10,000 total cells sampled, the difference between the distribution pre- and post- cell sampling decreases (Figure 3.2a). This difference, measured by changes in the MDI (Total Error), varied by type of distribution (Figure 3.2b). For example, the diversity of the uniform distribution was more difficult to fully sample than the exponential. This effect is also seen in the sampling needed to achieve half maximal species richness (as a percent of unique cells captured) (Figure 3.2c). While the uniform distribution requires ∼18% sampling to capture 50% of the diversity, the exponential and normal only require ∼10% to capture the same 50% of the diversity. Interestingly, we find that the extreme values of the distribution parameters were overall less significant than changing the distribution type. Overall, we find that the more polarized the distribution independent of its parameter values, the less sampling needed to minimize the error.

3.3.3 PCR of a monoclonal antibody generates a distribution of sequences

Next, we isolated the effect of PCR amplification on the various initial distributions. In order to examine the effects of PCR on the repertoire, we first considered the case of PCR on a population of antibodies with only one

unique identity, i.e. a monoclonal antibody repertoire. Any change to the repertoire introduced by the PCR would then be obvious as an expansion of the repertoire away from the single progenitor sequence.

We find that PCR of a single molecule of a single antibody identity can result in a large number of errors (Figure 3.3a, left). That is, only 95% of the final sequences are the original monoclonal sequence. Singleton errors comprise of only $\sim$0.02% of the overall counts, and can likely be attributed to errors arising in later cycles of PCR. In addition, since errors are generated at any cycle in the PCR process and subsequently amplified, large sections of the resulting tail are not just single errors, but rather amplified errors.

Moreover, we found that the number of starting molecules of a particular monoclonal antibody affects the PCR result. When a small number of starting molecules are used (N=1 and 10), the most abundant error (i.e. the second highest ranked sequence identity) is approximately two orders of magnitude less frequent than the correct sequence (the highest ranked sequence identity) (Figure 3.3a). In contrast, when a large number of starting molecules are used (N=1000), the correct sequence identity outnumbers the most abundant sequence error by four orders of magnitude.

Error due to the reverse transcriptase (RT) was hypothesized to contribute more to the error than DNA polymerase error, since RT enzymes typically exhibit an overall higher rate of error than DNA polymerase. Also, the RT treatment occurs in the first cycle, so errors are expected to be amplified more than errors occurring in later cycles. The resultant distribution when the

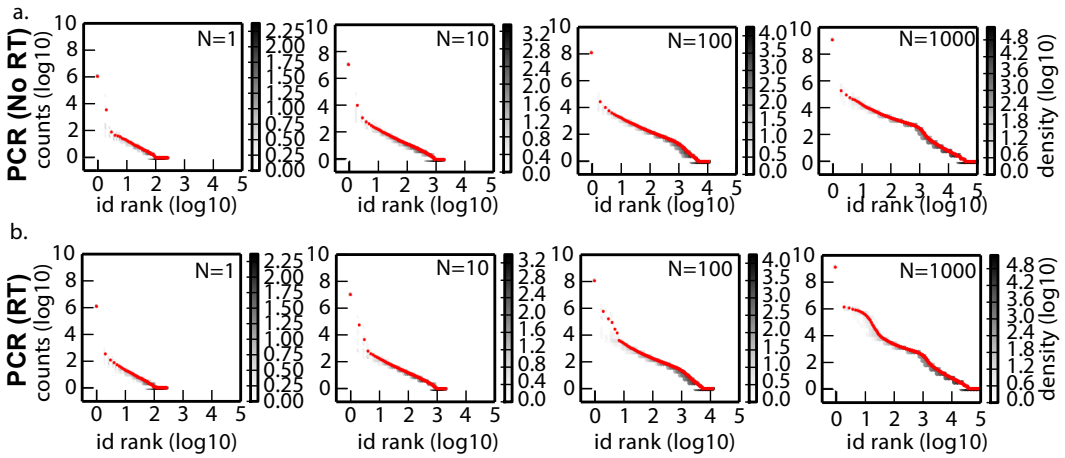


Figure 3.3: **Simulated distributions of PCR amplified monoclonal antibody** (a) Amplified sequence counts after 1 cycle of simulated reverse transcription with error rate of 0 (no error) and 19 cycles of simulated PCR with error rate of 4.4×10^{-7}, shown on log-log scale. (b) Amplified sequence counts after 1 cycle of simulated reverse transcription with error rate of 3.4×10^{-5} (Superscript III) and 19 cycles of simulated PCR with error rate of 4.4×10^{-7} (phusion), shown on log-log scale. N represents the number of molecules of the same identity that were amplified.

RT error is removed shows that the difference between counts of the original monoclonal sequence and the error is $\sim 10^{2.5}$. This is in contrast to Figure 3.3b, where the error rate of the RT is included in the first round of PCR, and there is a broadening of the left side shoulder, where that same difference is $\sim 10^{2}$.

Overall, removal of the RT step improves the discrimination between the first and second rank identities by half an order of magnitude. RT, while crucial to the HTS experimental workflow, makes it more difficult to use frequency as a measure of discrimination between the correct sequence and the subsequent tail of incorrect sequences.

When we ran multiple simulation trials (N=15) and visualized the variation in the resulting distributions, we found that the higher the number of starting molecules, the less scatter we observed (Figure 3.3, in grey). This is consistent with the hypothesis that early cycle mutations occurring in a small starting population of molecules will result in a large percentage of the incorrect sequences in the final distribution. In contrast, an early cycle mutation in a large starting population will be less likely to dominate the final distribution.

3.3.4 Sequencing increases the apparent number of unique sequences

In order to model the effects of different high-throughput sequencing platforms on the immune repertoire reconstruction, we took advantage of software for simulating the collection of high-throughput sequencing reads with realistic errors, as implemented in the ART simulation tools from the Na-

tional Institute of Environmental Health Sciences [30]. We hypothesized that sequencing more (i.e., higher number of reads) beyond a certain threshold of reads would result in a less accurate representation of the original initial antibody repertoire, due to additional sequencing error introduced by additional reads and lack of a reference sequence with which to correct those errors. Indeed, we find this to be the case (Figure 3.4), and in a simulation of the sequencing of a monoclonal repertoire, the apparent repertoire is considerably diversified even with a relatively small number of reads.

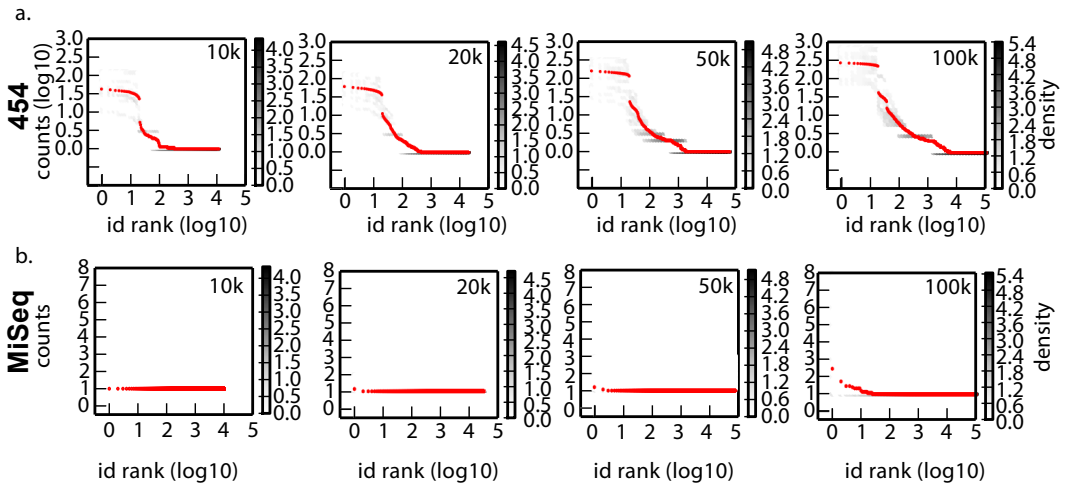


Figure 3.4: **Simulated high-throughput sequencing of a monoclonal antibody introduces additional error, broadening the apparent repertoire** (a) Sequence counts after generation of the indicated number of reads using the ART 454 read simulator. (b) Sequence counts after generation of the indicated number of reads using the ART MiSeq read simulator. Number of reads denoted in upper right corner of each plot and model assumes no error in prior PCR amplification (a-b).

We considered two alternate high-throughput sequencing platforms, the Roche 454 (the instrument used for many early repertoire sequencing projects

due to its ability to generate longer sequencing reads) and the newer Illumina MiSeq platform. We found that in both 454 simulations (Figure 43.4a) and MiSeq simulations (Figure 3.4b), increasing the number of read counts above 10,000 reads strictly increased the number of errors. When generating 10,000 reads, $\sim 1 \times 10^4$ unique sequence identities resulted and when generating 100,000 reads, $\sim 1 \times 10^5$ unique sequence identities resulted.

Furthermore, we found that the amount and shape of error generated varied by sequencing technology used. The MiSeq simulation shows a very tight distribution, but little discriminatory power between the correct monoclonal sequence and the remaining tail. Almost every sequence (>99%) is error. On the other hand, the 454 simulation shows a bipartite distribution. The left side is composed of mostly high frequency errors, where only the far left point is the correct sequence. The right side shows a fast decay of frequency of incorrect sequences. In both types of sequencing, generating more sequencing reads results in more incorrect sequences. Overall, sequencing adds a large amount of error to the process.

3.3.5 Clustering compensates for the PCR and sequencing inflation of unique sequence identifications

We have observed that each of the physical manipulations of the mRNA and DNA during repertoire sequencing serves to introduce errors: the copying of the immunoglobulin cDNA from expressed mRNAs, the amplification of the cDNAs by PCR, and the high-throughput sequencing of these molecules

into observed sequencing reads. The result is to generate clouds of sequences around the true sequences, related by minor (<5%) sequencing changes. Thus, clustering of the sequences potentially offers substantial error-correction and should in principle improve the reconstruction of the underlying repertoire. We therefore quantified the contribution of clustering to error correction of the simulated PCR and HTS data from earlier modules. Note that in contrast to these earlier modules, which computationally simulated experimental methods, this part of the simulation replicates the actual informatic methods used to analyze data from immune repertoire HTS experiments.

We find that clustering at a threshold of at least 98% amino acid identity largely removes the effect of the PCR error (Figure 3.5a), whereas clustering at a threshold of at least 90% identity is needed to remove high-throughput sequencing error (Figure 3.5b), indicating that sequencing is a larger source of error compared to PCR. Interestingly, clustering thresholds necessary to recapitulate the original distribution are similar whether or not RT is included, indicating that while RT introduces abundant high rank errors, they are correctable via clustering (Figure 3.5a). In addition, while 454 sequencing results in a smaller number of total errors than MiSeq sequencing, a clustering threshold of 90% is equally effective at reducing the error in both types of high throughput sequencing (Figure 3.5b). Our conclusion is that clustering is critical for a more faithful reconstruction of the antibody repertoire. However, it is important to note that clustering only compensates for errors introduced via physical copying and sequencing of the mRNA/DNA, and not due to cell

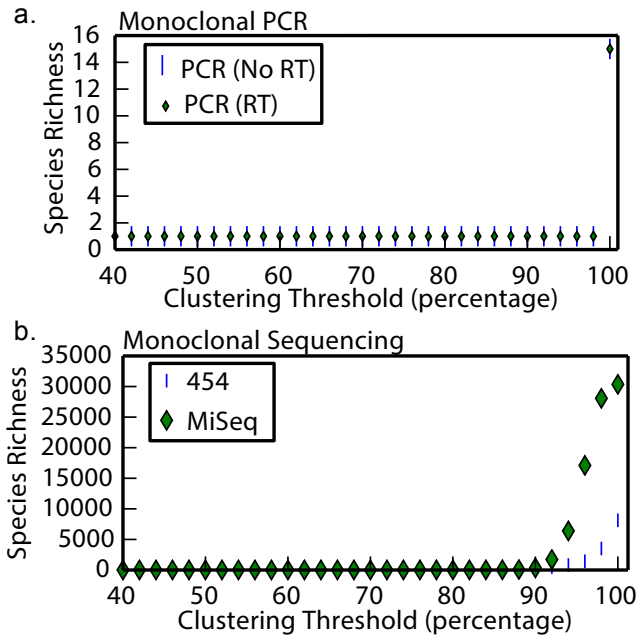


Figure 3.5: **Sequence clustering largely corrects the PCR and sequencing-induced errors in a monoclonal antibody repertoire** (a-b) Species richness, defined as the number of unique sequences represented, calculated from distributions obtained after clustering (40%-100% thresholding). Clustering was applied to the results of monoclonal PCR (shown in Figure 3.3) either with (green diamonds) or without (blue vertical bars) RT error in (a). Clustering was applied to the results of sequencing (Figure 3.4) with either the 454 simulator (blue vertical bars) or the MiSeq simulator (green diamonds).

sampling.

3.3.6 The full model recapitulates the empirical distribution

We simulated the entire pipeline shown in Figure 3.1b and found that in order to recover the original species richness (N=3,000), a ∼75% clustering threshold is needed (Figure 3.6a). In addition, we find that the final empirical distributions, which result from simulations using three distinct theoretical biological distributions, are indistinguishable from each other (Figure 3.6b). Specifically, each empirical distribution is Zipfian and the species richness is higher than the species richness of the theoretical biological distribution. Also, the counts are continuous and there is no threshold count value that intuitively separates the high count from the low count identifications.

3.4 Discussion

In this study, we simulated the HTS immunoglobulin repertoire workflow from sample collection through data analysis. We isolated individual modules of error including cell sampling, PCR, and HTS. The amount of cell sampling required to sample half of the diversity is relatively small (∼14%), but would still be prohibitive to perform in humans. Within these modules, we find that PCR and sequencing contribute the most to the overall error in the process. We also find that aggressive clustering to be a simple error correction method for restoring the empirical distribution to the species richness of the original theoretical distribution. Strikingly, we found that even

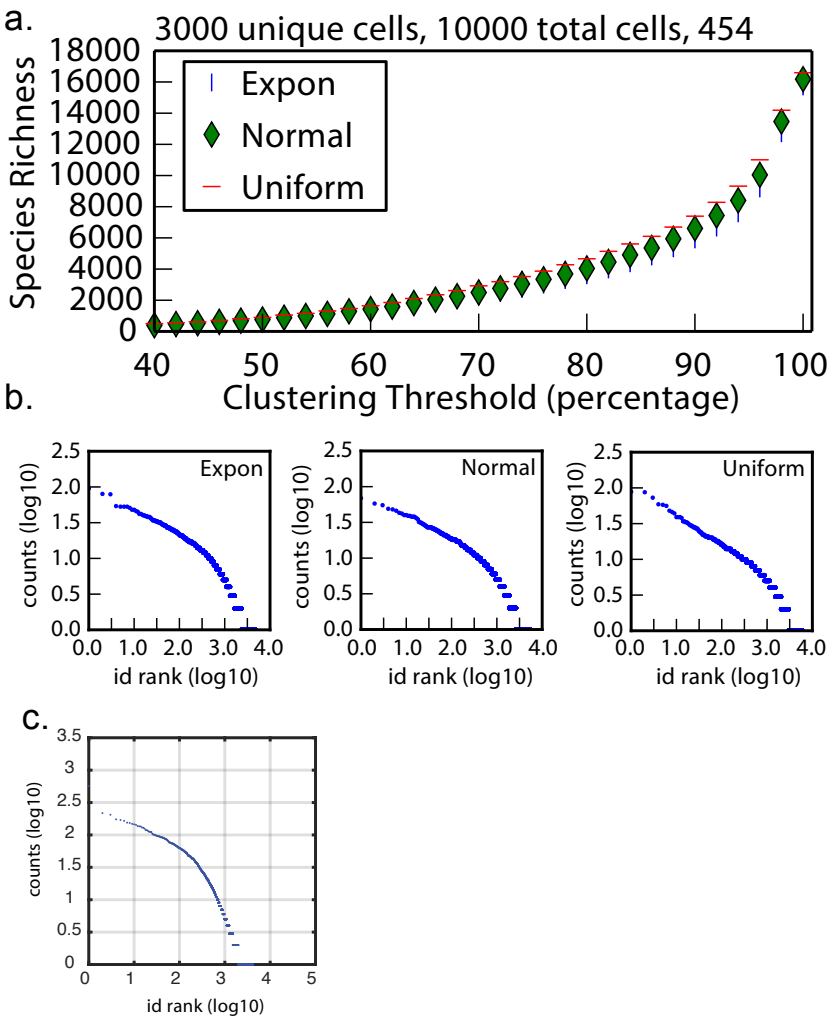


Figure 3.6: **Antibody repertoire distributions resulting from the complete simulation pipeline resemble empirical distributions** (a) Species richness resulting from simulation of cell sampling, PCR amplification, 454 sequencing, and clustering from 40%-100% thresholding levels. Initial distributions were either exponential (blue vertical bars), normal (green diamonds) or uniform (red horizontal dashes). (b) Clonal cluster count distributions from complete simulation (using 86% clustering threshold). Initial distribution types are indicated in the top right corner. Clonal cluster IDs are sorted in descending order and plotted on a log-log scale. (c) Sequence count distribution from experimental data [64].

a single molecule, when processed through our pipeline, can generate a full spectrum of error that is comparable to empirical distributions from experimental data. It has been proposed that the true biological distribution could be deduced from the empirical distribution [49], and our model attempted to provide a computational algorithm for doing so. However, this remains an unsolved challenge, since using different theoretical biological distributions as input resulted in the same empirical distributions that were indistinguishable from one another. Specifically, the true biological distribution may be normal, exponential, or even uniform and not necessarily Zipfian, as the empirical distributions might suggest. This has implications for our understanding of the diversity of the immune system; the true diversity and shape of the distribution may deviate significantly from what the experimental data indicates. In this way, the HTS immunoglobulin repertoire workflow may be a many-to-one function, where the empirical distribution may not have an inverse function that can be used to determine the true biological distribution.

It is important to note that we did not attempt to fit the parameters of the simulations to optimize the match between the simulated and empirical distributions. Rather, we used established error rates and estimates of the sampling regime that best matched our experimental setup. The resulting agreement between the simulated and empirical repertoire distributions emerges without further optimization, and suggests that the simulation is correctly recapitulating the major sources of error in the real pipeline.

Based on our simulation results, we offer several recommendations for

HTS immune repertoire data generation and interpretation. First, due to the observation that every theoretical biological distribution resulted in highly similar final empirical distributions, we recommend that no claims on the true biological distribution be made based on the empirical distribution alone. Second, our results indicate that PCR and HTS are two major sources of error. Specifically, given the significant amount of error generated by the first cycle in PCR, we recommend using a reverse transcriptase with lower error rates such as RTX [18]. In addition, because PCR is an exponential process, minimizing the number of PCR cycles will greatly reduce error generation. Finally, our simulation suggests that aggressive clustering, as we showed using a 75% identity threshold, may be important for reducing amplification and sequencing errors to restore the original species richness.

Future experiments might involve deliberately modifying the experimental protocol to test these assumptions (for example, by using a error correcting reverse transcriptase [18]). Alternatively, sweeping the parameter choices of the simulation to find a better fit to the experimental distribution could be used as a means of better estimating the true error rates.

Our model can be tailored to different experimental parameters as needed, by swapping in different PCR emulators, sequence generators, and clustering algorithms. In addition, our model is versatile enough to allow for simulation of HTS of systems other than the immune repertoire of humans, such as immune systems of zebrafish or metagenomics in ecological niches. Common methods of error correction, such as sequence barcoding, could also

be included as additional modules to increase the utility of the simulation.

3.5 Materials and methods

A sequence database was generated from previously published data [64]. The sequences were annotated with VDJFasta, and filtered for valid V and J gene calls. For the PCR simulation, sequences were randomly selected from this database and mutated probabilistically, consistent with the error rate of the theoretical polymerase at each cycle of PCR. The mutated sequence were then used for subsequent analysis in downstream modules. Clustering was performed across full length amino acid sequences.

We chose to focus on a small number of widely prevalent distributions in nature that vary dramatically in their levels of polarization; namely, the uniform, normal, and exponential. The uniform distribution is a simple and extreme distribution in that it shows no polarization; there is no difference in frequency to any individual identity. It can be visualized as a horizontal line. It is defined entirely by two parameters: the total number of unique identities and the total number of members. We also chose a highly polarized distribution, the exponential distribution, where a small number of individual identities represent a disproportionate fraction of the total number of members. In addition to the parameters required in the uniform distribution, the exponential distribution requires one additional parameter, the rate parameter, represented by λ. Variation of λ can change the degree of polarization of the exponential distribution. Finally, we chose a distribution of intermediate

polarization, the normal distribution. This distribution is highly prevalent and appears in fields as disparate as social science and theoretical physics. The normal distribution is defined by two parameters in addition to the two parameters that define the uniform distribution: the mean, μ, and the variance, σ. The mean is the value where the most highly frequent identities appear and the variance describes how far away identities are from the mean.

We chose values of 0.1, 0.3, and 0.5 for lambda to cover a wide range of polarization in the exponential distribution. The mean in the normal distribution does not affect the degree of polarization; therefore, we fixed it at 10 for convenience, and then chose the values of 0.2, 0.5, 1, and 5 for the variance to cover a wide range of polarization for the normal distribution. Ultimately, the specific parameter choices for any given distribution was less important than the type of distribution chosen. The distributions maintained the same relative degree of polarization between distributions (i.e., exponentials were always more polarized than normal, which were always more polarized than uniform). This change in polarization was less extreme within a distribution as compared to between distributions. We also found that the ratio of unique sequences to total cell number was more important than the absolute values. We chose values that created sufficient diversity consistent with current experimental procedures while maintaining reasonable computational demands. For the full simulation, we ran the full simulation assuming 10,000 total cells and 3,000 unique cells. We sampled separately 2,000 cells from a uniform distribution, an exponential with the rate parameter, λ=0.1, and a normal

distribution with the variance, σ=1. We simulated 20,000 454 reads.

Our computational models were programmed using Python. Visualizations were all made using Python and matplotlib. High-throughput sequencing simulations were performed using ART [30]. FLASH [42] was used to combine paired-end sequences, when necessary. VDJFasta [25] was used to annotate sequences. Reference genes were acquired from the international ImMunoGeneTics (IMGT) [1] database. Clustering was done using CDHit [22].

3.6 Acknowledgements

We thank Dr. Weichun Huang for help adapting the 454 simulator for the purposes of our simulation, and Drs. Gregory Ippolito and George Georgiou for the initial stimulus of this project and many helpful discussions. E.M.M. acknowledges grant support from DTRA, NIH, NSF, and the Welch Foundation (F1515).

3.7 Author contributions

EMM and GCW designed the simulation. GCW implemented the simulation and ran the experiments. GCW wrote the manuscript under the supervision of EMM. GCW and EMM edited the manuscript.

Chapter 4

Conclusion

The immune repertoire is intricate and vast. A better understanding of immune cells at both the tissue and molecular level teaches us how we continue to survive in an environment that can reproduce, grow, and adapt faster than we can. The antibody repertoire is one particularly large part of the immune repertoire. The work presented here gives insight into the bone marrow plasma cell repertoire, a specific aspect of the human antibody repertoire. More broadly, we also characterized a method used in the study of the immune repertoire, high-throughput sequencing (HTS). Future work will improve the quality of data interpretation and the depth of biological insight. Here, I propose future simulations and experiments that could address some of the remaining outstanding questions in the study of the immune repertoire.

My work demonstrates that high-throughput sequencing is a powerful, if imperfect tool for studying the antibody repertoire. I evaluated and quantified one particular clustering method for informatic correction of amplification and sequencing errors. A variety of different clustering algorithms are used commonly in error correction for HTS, with different parameters and use cases (e.g. phylogenetic clustering). Many of these algorithms are op-

timized for speed, but accuracy is an important and previously intractable optimization parameter. Future simulations can build on the one presented here in order to evaluate the accuracy of error correction methods and resolve conflicting hypotheses on proper data analysis and interpretation. Another technical challenge in immune repertoire analysis is optimizing diversity index choice to provide the most comprehensive and accurate view on the true diversity. There are many methods of measuring diversity, including Simpson, Shannon, and Berger-Parker. The simulation presented in Chapter 3 may provide a framework that enables the comparison of various diversity indices. By controlling the inputs and parameters of the simulation, particularly the true biological distribution, and observing the resulting outputs, it will be easier to understand the benefits and limitations of each diversity metric. In addition to the technical questions discussed above, the nascent immune repertoire field continues to pursue unanswered biological questions.

Beyond technical methods of ensuring accurate data generation, the antibody repertoire is a complex mix of diverse sequences and functions that elude simple interpretation. At the level of individual sequences—even corrected ones—there is too much noise for biological interpretation. There is a need to group these sequences with the intent of biological interpretation. One potential solution would be to look at the evolutionary relationships of the sequences by performing ancestral sequence reconstruction by phylogenetic analysis. This is possible because antibody sequences are known to undergo sequential evolutionary steps involving somatic hypermutation to increase bind-

ing affinity without changing antigen specificity. Phylogenetic analysis, therefore, could provide a method of classification for antibodies, which would group individuals with similar, but not identical, immune repertoires and identify immune signatures. These signatures can then be used in a wide variety of research and clinical applications, such as targeted therapy development for personalized medicine.

In this work, I observed that the immune repertoire is remarkably stable across time; however, within a single timepoint, the biological resolution of the immune repertoire is still fairly limited. While immune repertoire composition on the sequence level was observed, the cellular level information was not attainable. Concretely, the heterogeneity in gene expression among a population of single plasma cells remains poorly understood. Early studies suggested that plasma cells only produce one type of antibody [58]. From studies of B cell derived cell lines, the amount of antibody produced per cell was observed to vary considerably [8]. However, it is not known how much of this variation is due to the B cell developmental stage, how much is the consequence of active regulation, and how much is due to cellular noise. Deep sequencing of plasma cells in bulk like those used in this work cannot distinguish between the possibility of one cell producing large amounts of immunoglobulin and many cells producing smaller amounts of immunoglobulin, nor can deep sequencing identify sources of noise that are hidden by bulk sampling. Emerging single cell analysis platforms will enable the study of the heterogeneity of immune cells.

The primary function of the antibody is to recognize, bind, and neu-

tralize potentially dangerous assaults on the host. The major force shaping the immune repertoire is therefore the antigen repertoire, which high-throughput sequencing, even at the single cell level, cannot capture. This is because the possible repertoire of antigens goes beyond nucleic acids and includes other macromolecules like carbohydrates and proteins. The most promising technology for pursuing the antigen repertoire is mass spectrometry. This proteomic technology has already been used as a powerful antibody discovery tool. However, even this technology would only have the ability to look at the protein antigen repertoire. The full extent of the antigen repertoire remains to be discovered.

Strikingly, all these questions and areas of research can be posed and pursued in different immunological compartments, cell types, and animal models. The field of human repertoire analysis remains vibrant, accelerated by current technologies, and holds tremendous progress as it anticipates future breakthroughs. The immune system is an exquisite system that protects the host from danger. It is finely tuned to recognize self from non-self, danger from benign. As we learn more about the immune systems intricacies, we advance toward the capability to fortify and ameliorate it when that delicate balance is lost.

Appendices

Appendix A

A Census of Human Soluble Protein Complexes[1,2]

[1]Havugimana, PC [*et al.*, including Wu, GC]. A Census of Human Soluble Protein Complexes. Cell, 150(5):10681081, August 2012.

[2]In the following work, I performed cloning and AP/MS experiments to validate protein-protein interactions of predicted associations. I also edited and reviewed the manuscript prior to publication and approved the final version.

Cell

Resource

A Census of Human Soluble Protein Complexes

Pierre C. Havugimana,[1,2,8] G. Traver Hart,[1,2,8] Tamás Nepusz,[4,8] Haixuan Yang,[4,8] Andrei L. Turinsky,[5] Zhihua Li,[6] Peggy I. Wang,[6] Daniel R. Boutz,[6] Vincent Fong,[1] Sadhna Phanse,[1] Mohan Babu,[1] Stephanie A. Craig,[6] Pingzhao Hu,[1] Cuihong Wan,[1] James Vlasblom,[2,5] Vaqaar-un-Nisa Dar,[7] Alexandr Bezginov,[7] Gregory W. Clark,[7] Gabriel C. Wu,[6] Shoshana J. Wodak,[2,3,5] Elisabeth R.M. Tillier,[7] Alberto Paccanaro,[4,*] Edward M. Marcotte,[6,*] and Andrew Emili[1,2,*]

[1]Banting and Best Department of Medical Research, Donnelly Centre for Cellular and Biomolecular Research

[2]Department of Molecular Genetics, Medical Sciences Building

[3]Department of Biochemistry, Medical Sciences Building

University of Toronto, Toronto, Ontario M5S 3E1, Canada

[4]Department of Computer Science, Royal Holloway, University of London, Egham TW20 0EX, UK

[5]Hospital for Sick Children, 555 University Avenue, Toronto, Ontario M5G 1X8, Canada

[6]Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, Department of Chemistry and Biochemistry, University of Texas at Austin, Austin, TX 78712, USA

[7]Campbell Family Institute for Cancer Research, Ontario Cancer Institute, University Health Network, University of Toronto, Toronto, Ontario M5G 1L7, Canada

[8]These authors contributed equally to this work

*Correspondence: alberto.paccanaro@cs.rhul.ac.uk (A.P.), marcotte@icmb.utexas.edu (E.M.M.), andrew.emili@utoronto.ca (A.E.)

http://dx.doi.org/10.1016/j.cell.2012.08.011

SUMMARY

Cellular processes often depend on stable physical associations between proteins. Despite recent progress, knowledge of the composition of human protein complexes remains limited. To close this gap, we applied an integrative global proteomic profiling approach, based on chromatographic separation of cultured human cell extracts into more than one thousand biochemical fractions that were subsequently analyzed by quantitative tandem mass spectrometry, to systematically identify a network of 13,993 high-confidence physical interactions among 3,006 stably associated soluble human proteins. Most of the 622 putative protein complexes we report are linked to core biological processes and encompass both candidate disease genes and unannotated proteins to inform on mechanism. Strikingly, whereas larger multiprotein assemblies tend to be more extensively annotated and evolutionarily conserved, human protein complexes with five or fewer subunits are far more likely to be functionally unannotated or restricted to vertebrates, suggesting more recent functional innovations.

INTRODUCTION

Protein complexes are stable macromolecular assemblies that perform many of the diverse biochemical activities essential to cell homeostasis, growth, and proliferation. Comprehensive characterization of the composition of multiprotein complexes in the subcellular compartments of model organisms like yeast, fly, worm, and bacteria have provided critical mechanistic insights into the global modular organization of conserved biological systems (Hartwell et al., 1999), accelerated functional annotation of uncharacterized proteins via guilt by association (Hu et al., 2009; Oliver, 2000), and facilitated understanding of both evolutionarily conserved and disease-related pathways (Vidal et al., 2011). How the ~20,000 or so proteins encoded by the human genome are partitioned into heteromeric "protein machines" remains an important but elusive research question, however, as less than one-fifth of all predicted human open reading frames are currently annotated as encoding subunits of protein complexes in public curation databases (Ruepp et al., 2010).

Loss-of-function mutations in genes encoding the subunits of protein complexes typically give rise to similar phenotypes or, through genetic interaction, amplify the phenotypic effects of other alleles in functionally linked sets of genes. Identifying the membership of protein complexes, therefore, addresses a crucial layer in the hierarchical functional organization of biological systems that links the core biochemistry of a functioning cell to the general physiology of an organism and is fundamental to deciphering the relationship between genotype and phenotype. Although bioinformatics analyses have been used to predict evolutionarily conserved human protein-protein interactions (PPIs) on a large scale (Ramani et al., 2008; Rhodes et al., 2005), most of these associations remain to be verified experimentally.

Affinity purification (AP) of tagged exogenous proteins coupled with tandem mass spectrometry (MS) is an effective method for isolating and characterizing the composition of stably associated human proteins in experiments ranging from dozens to hundreds of different "baits" (Behrends et al., 2010; Bouwmeester et al., 2004; Ewing et al., 2007; Hutchins et al., 2010; Jeronimo et al., 2007; Mak et al., 2010; Sardiu et al., 2008; Sowa et al., 2009). Likewise, immunoprecipitation can be used

to systematically isolate endogenous human protein complexes from human cell lines (Malovannaya et al., 2011). Nevertheless, the limited availability of high-quality antibodies or sequence-verified complementary DNA (cDNA) clones suitable for targeted protein complex enrichment precludes scale-up required for the unbiased assessment of the molecular association networks underlying human cells. Hence, despite considerable successes in the comprehensive identification of protein complexes in model organisms (Butland et al., 2005; Gavin et al., 2002, 2006; Guruharsha et al., 2011; Ho et al., 2002; Hu et al., 2009; Krogan et al., 2006; Kühner et al., 2009), clone-based protein purification techniques remain challenging for proteome-scale studies of physical interaction networks in mammalian cells. Conversely, although traditionally used to isolate discrete complexes with specific assayable biochemical properties (e.g., enzymatic activity), classical biochemical fractionation procedures have been used to resolve biological mixtures as a means of ascertaining the collective composition of human protein complexes present in certain subcellular compartments (Ramani et al., 2008; Wessels et al., 2009).

Here, we have combined extensive, scaled-up biochemical fractionation with in-depth, quantitative mass spectrometric profiling and stringent computational filtering to resolve and identify endogenous, soluble, stably associated human protein complexes present in cytoplasmic and nuclear extracts generated from cultured cells. Although the resulting reconstructed high-quality physical interaction network shows strong overlap with existing curated and experimentally derived sets of annotated protein complexes, it contains many predicted subunits and previously unreported complexes with specific functional, evolutionary, and disease-related biological attributes. To our knowledge, this resource represents the largest experimentally derived catalog to date of human protein complexes from cell culture, measured using a single standardized assay, and a reliable first draft reference of the basic physical wiring diagram of a human cell.

RESULTS

High-Throughput Complex Fractionation and Detection by Tandem Mass Spectrometry

To isolate human protein complexes in a sensitive and unbiased manner, we subjected cytoplasmic and nuclear soluble protein extracts isolated from human HeLa S3 and HEK293 cells grown as suspension and adherent cultures, respectively, to extensive complementary biochemical fractionation procedures. These two widely studied laboratory cell lines have been used as models of human cell biology for many decades (Graham et al., 1977; Masters, 2002), providing a rich biological context for interpreting the resulting proteomic data. Stably interacting proteins that cofractionated together were identified subsequently by nanoflow liquid chromatography-tandem mass spectrometry (LC-MS/MS). We optimized our entire experimental pipeline, illustrated schematically in Figure 1A, by using a multipronged strategy to minimize two major confounding issues: limited dynamic range (i.e., preferential detection of high-abundance components) and "chance" coelution (i.e., cofractionation of functionally unrelated proteins).

To address the former concern, we performed extremely deep biochemical fractionations by employing multiple orthogonal separation techniques to better resolve distinct protein complexes. As a primary separation technique, we employed nondenaturing high-performance multibed ion exchange chromatography (IEX-HPLC) by using four different empirically optimized analytical column combinations (see Experimental Procedures) and shallow salt gradients unlikely to perturb nonionic protein associations (Havugimana et al., 2007). In parallel, we applied complementary sucrose gradient centrifugation and isoelectric focusing technologies to capture salt-sensitive protein assemblies. In total, we collected 1,163 different fractions in a total of eight nuclear and five cytosolic extract fractionation experiments (for details see Table S1 available online), which were each subjected to label-free shotgun sequencing (duplicate LC-MS/MS analyses) using highly sensitive ion trap-based mass spectrometers (see Experimental Procedures).

We identified 5,584 distinct human proteins (Figure 1C; estimated theoretical false discovery rate of 1% at both the protein and peptide levels based on a statistical model [Kislinger et al., 2003]; see Experimental Procedures for details). Despite the underrepresentation of membrane proteins in the starting cell extracts, this coverage encompasses about half of the experimentally verified human proteome (Figure S1B) (Nagaraj et al., 2011). This included 989 proteins detected exclusively in nuclear fractions (of which 376 were annotated transcription or chromatin-related factors) and 1,006 with links to human disease (e.g., annotated in a public database like OMIM). Only 1,632 (29%) of the identified proteins had biochemical annotations as subunits of previously reported protein complexes (corresponding to 64% of all existing human protein entries) in the CORUM curation database (Figure S1C; Ruepp et al., 2010). Due to the extensive fractionation, we observed minimal bias in terms of protein abundance beyond that reported for previously annotated complexes or the experimentally defined human proteome (Figure 1D).

Next, to minimize the possibility of chance coelution, rather than simply identifying the proteins present in each fraction, we quantified variation in protein abundance based on the observed patterns of spectral counts recorded across all of the collected fractions to determine the extent to which pairs of proteins coeluted. As shown in Figure 1B, these experimental profiles were highly reproducible (i.e., average Spearman rank correlation coefficients >80% between replicate experiments; Figure S2), even using alternate methods of mass spectrometric quantification (i.e., extracted MS1 peak intensities were largely consistent with spectral counting; Figure S2D). To objectively evaluate the biochemical data, we calculated a stringent summary statistic, termed the coapex score, for each pair of proteins identified by LC-MS/MS by determining the number of fractionation experiments in which the proteins showed maximum (modal) abundance in the same exact peak fraction.

To assess the effectiveness of our cofractionation approach, we performed an initial validation by examining the coelution profiles and coapex scores obtained for a reference set of 20 well-known human protein complexes reported in CORUM. As illustrated by the representative HeLa nuclear extract IEX-HPLC profiles shown in Figure 1B, the subunits of these

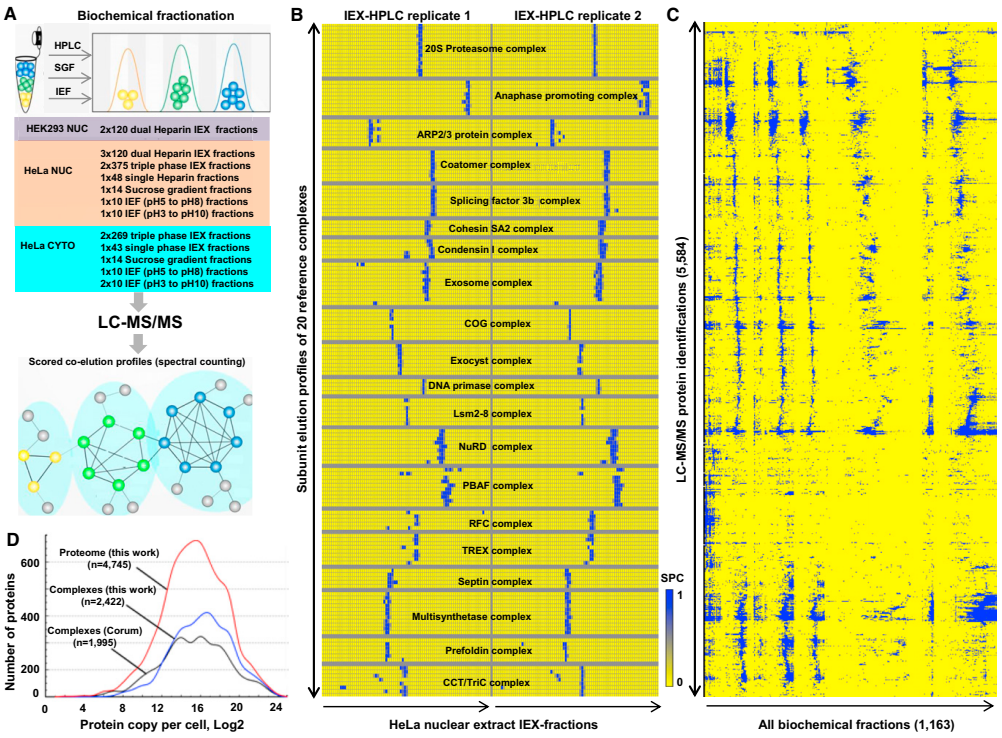


Figure 1. Integrative Cofractionation Strategy Used to Identify Human Soluble Protein Complexes

(A) Cell extracts were extensively fractionated using different biochemical techniques (IEX, ion exchange chromatography; IEF, isoelectric focusing; SGF, sucrose density gradient centrifugation). Coeluting proteins were identified by mass spectrometry, and a coelution network was generated by calculating profile similarity (see Extended Experimental Procedures).

(B) Cofractionation (IEX-HPLC) profiles of annotated subunits of 20 representative human protein complexes from HeLa nuclear extract. Shading indicates normalized spectral counts (SPC). Peak apex and adjacent peaks are shown.

(C) Hierarchical clustering of 5,584 proteins identified by LC-MS/MS.

(D) Protein abundance levels corresponding to components of our identified coeluting proteins (red line), reconstructed complexes (blue), or annotated CORUM complexes (black) estimated from the reported HeLa proteome (Nagaraj et al., 2011).

See also Figure S1 and Table S1.

complexes typically coeluted in the same biochemical fractions. Of the 155 components detected by mass spectrometry, most (85%; 499/585) of the detected subunit pairs of the reference complexes had high coapex similarity scores (i.e., coeluted together in at least two or more experiments), validating the overall efficacy of the fractionation procedures we used to isolate native protein complexes and the general correctness of the protein identification and quantification pipeline.

Reconstruction of a High-Confidence Cocomplex Interaction Network

Despite the consistency in coelution of annotated complex members, certain functionally distinct complexes occasionally exhibited overlapping chromatographic elution profiles (e.g., splicing factor 3b and Coatomer complexes; Figure 2A), presenting a potential source of spurious interactions. Although this artifact was minimized to a certain degree by performing multiple independent fractionation experiments, we used an integrative computational approach to further improve deconvolution (Figure 2B). Because physically interacting cocomplexed proteins often perform related biological functions (Alberts, 1998) and are often evolutionarily coconserved (Hartwell et al., 1999), we devised a machine learning procedure (Figure 2B; see Experimental Procedures for details) to score and select higher-confidence physical interactions based on both the experimentally measured coelution profiles and the existence of additional

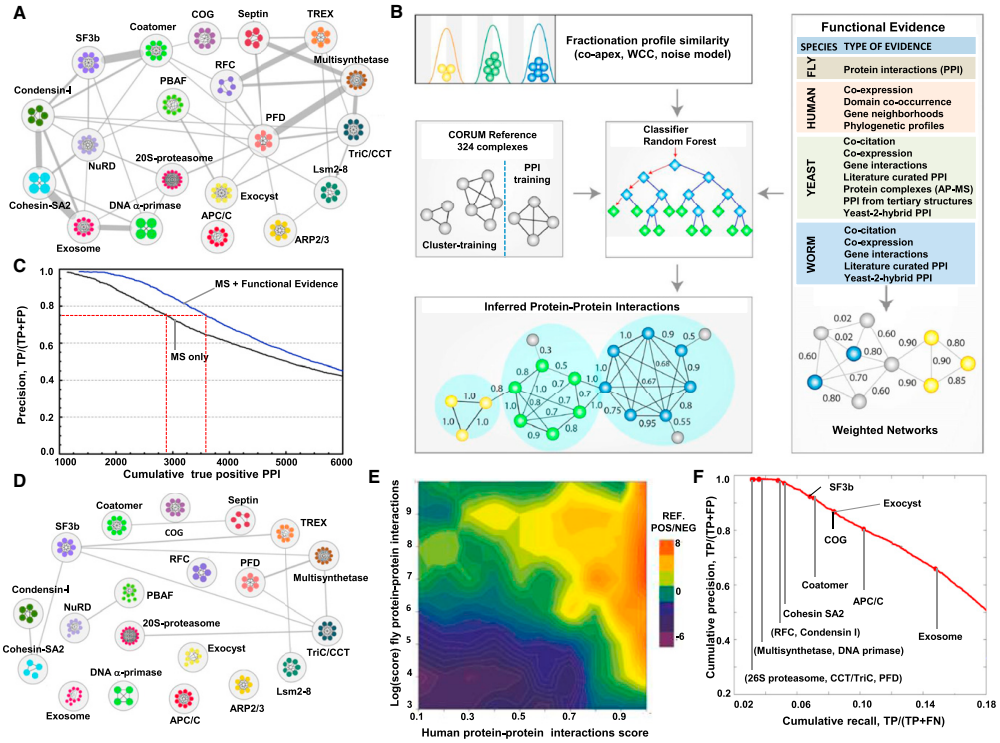


Figure 2. Denoising the Biochemical Coelution Network and Generation of High-Confidence Physical Interactions

(A) Biochemical cofractionation network of 20 reference complexes with coelution coapex scores ≥ 2. Nodes represent protein subunits (colors reflect complex membership), whereas edges represent interactions (thickness proportional to the number of shared coapexes).

(B) The biochemical data were combined with weighted functional association evidence by using a Random Forest classifier and a training set of reference complexes (CORUM) to filter out spurious connections and to infer a high-confidence interactome. The PPI and predicted clusters were evaluated with independent functional criteria to ensure high quality. Arrows represent data flow, blue diamonds are attributes in the decision tree vector, and green diamonds (leafs) are the final result (positive or negative).

(C) Cumulative precision-prediction rank curves for the LC-MS/MS data alone and after integration with genomic evidence. Incorporation of the functional evidence increased both precision (reduced false positives) and recall (more true positives).

(D) Network of 20 reference complexes after filtering with functional evidence.

(E) Overall correlation (Spearman r = 0.40; n = 11,675) of our scored human PPI with corresponding interaction scores reported for orthologous fly PPI from which validated, high-confidence complexes were derived (Guruharsha et al., 2011). Heatmap shows prediction accuracy (log ratio of CORUM reference positives to negatives), with high-scoring pairs in both studies highly enriched for positives.

(F) Precision-recall curve showing performance obtained after denoising reconstructing withheld reference CORUM complexes highlighted by red dots at the threshold at which half of the protein pairs per complex are recovered.

See also Figure S5 and Table S2.

supporting functional association evidence inferred from correlated evolutionary rates (Tillier and Charlebois, 2009) and functional genomics data sets compiled for *H. sapiens*, *S. cerevisiae*, *D. melanogaster*, and *C. elegans* (see Table S6 for details).

First, for each of the 13 fractionation experiments, we calculated correlation measures between all possible pairs of proteins to capture their tendency to coelute. In addition to the coapex summary statistic, to account for mass spectrometry sampling error, we devised a weighted cross-correlation function to account for slight variation in the protein profiles measured in each experiment. To account for low spectral values, we also generated a Poisson noise model before calculating Pearson correlation scores, deeming the coelution profiles of protein pairs measured with low spectral counts as less predictive of genuine physical interactions (Figure S5). Only protein pairs

with a correlation score of at least 0.5 by at least one of these measures in one or more experiments were considered for further analysis, reducing the total number of pairs from over 15 million initially to the roughly 800,000 pairs with reasonable biochemical evidence.

To improve the assignment of interaction probabilities, we also exploited the predictive power of correlated protein evolutionary rates (Tillier and Charlebois, 2009), messenger RNA (mRNA) coexpression, and domain co-occurrence and, via orthology, fly protein-protein interactions (based on binary yeast two-hybrid assay studies) and extensive physical and functional associations reported previously for yeast and worm (see Experimental Procedures) (Lee et al., 2011). The discriminatory power of the procedure was further improved by penalizing those interactions that lacked independent supporting evidence—and that were thus more likely to correspond to cases of ''chance'' coelution—by integrating evidence from these functional association data (Figure 2B). A feature selection algorithm was used to select the most informative data sets (Table S2) in addition to the biochemical correlation scores, and the resulting features were used to estimate the probability of interaction to protein pairs using a cross-validated random forest classifier.

For training, we used the CORUM curated set of human protein complexes as our base reference, filtered for those complexes annotated as having been observed by biochemical methods. As many CORUM complexes are highly overlapping due to redundancy in existing annotations, we combined complexes sharing subunits (Simpson coefficient >0.5 between complexes). We used half of the resulting 324 nonredundant reference complexes (Table S3) as the training set for cocomplex probability prediction, defining gold standard positive interactions as pairs of proteins in the same complex and inferring gold standard negatives between proteins in different complexes. (The other half of the reference complexes was withheld for subsequent use as an independent training set for cluster optimization, as described below.)

Although the biochemical data were a prerequisite for scoring, the performance curves shown in Figure 2C indicate that the inclusion of the additional functional genomic information substantially increased recall at the same level of precision compared to classifiers based on the profiling data alone. Moreover, the integration of this additional supporting functional evidence removed the bulk of spurious, intercomplex interactions (Figure 2D). Another advantage of our bioinformatic pipeline is that the results of the feature selection algorithm (Table S2) can be explored to examine the impact of each data set. For example, we find generally that sets of smaller biochemical fractionations using different separation techniques, although individually yielding a higher PPI false discovery rate, collectively provided more information on protein complex composition than deeper fractionations using a single separation method.

As an alternate measure of reliability, we compared our scored human protein interactions to a recently reported network of *Drosophila* cocomplex protein interactions (Guruharsha et al., 2011), which had not been used for building the classifier. Strikingly, despite using vastly different experimental methods and scoring schemes, we observed a remarkably good overall correlation (Spearman r = 0.40; n = 11,675 orthologs mapped using Inparanoid). Even after removing interactions supported by alternate *Drosophila* data, high-scoring fly pairs matched high-scoring pairs in our analysis and were strongly enriched for reference-positive cocomplex members (Figure 2E).

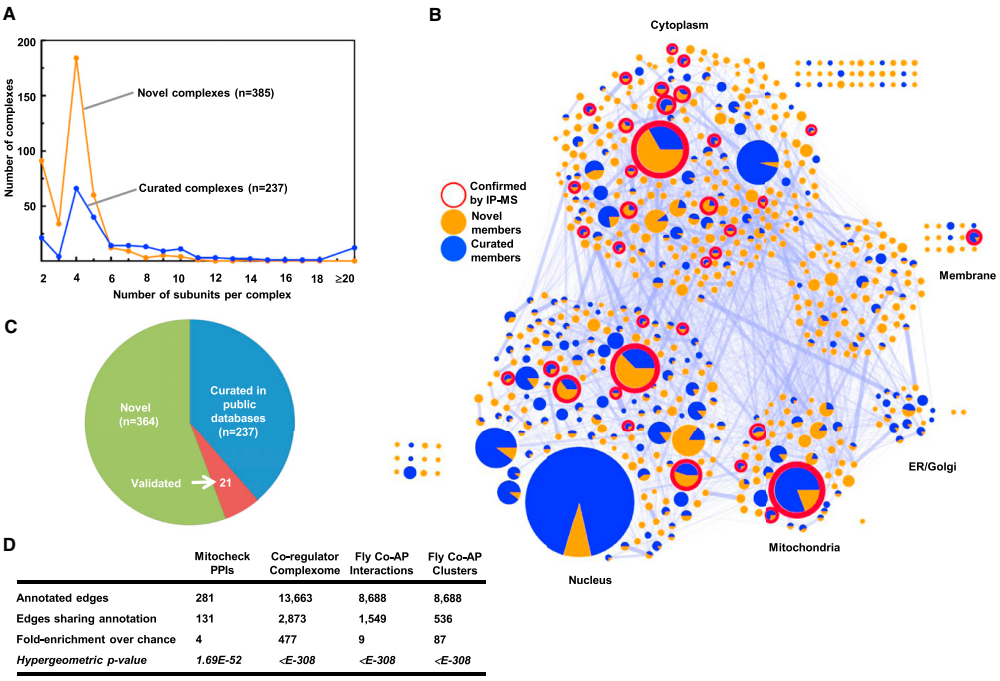
Finally, in order to remove any remaining false positive interactions, we further denoised our cocomplex data set by pruning loosely connected interactions using a computational diffusion procedure calibrated by protein colocalization semantic similarity scores (Pesquita et al., 2009; Yang et al., 2012) to enforce local network topologies more consistent with annotated complexes from the withheld portion of the reference Corum complexes (see Experimental Procedures). Benchmark precision and recall versus the holdout set of known reference complexes (Figure 2F) were significantly higher than those reported for a smaller, recently published set of affinity-purified human protein complexes (Hutchins et al., 2010), validating the reliability of our scoring procedure.

Applying a PPI score threshold of 0.75, which corresponds to an estimated false discovery rate of 21.5% (i.e., well below the ~40% reported for AP/MS-based analyses of protein complexes in model organisms [Gavin et al., 2006; Krogan et al., 2006; Kühner et al., 2009]), we thus derived a high-confidence set of 13,993 cocomplex interactions among 3,006 unique human proteins (Table S2), most of which (8,691 PPI) have not been reported before (i.e., are not publicly annotated). It is worth reiterating that all of these physical interactions were directly supported by the experimental biochemical cofractionation data; the addition of functional data and denoising served only to flag candidates lacking either functional support or topological support within the network (Table S2). The interaction probability scores may be underestimated, however, because the reference ''gold standards'' used for learning are imperfect (Jansen and Gerstein, 2004).

Construction and Validation of Protein Complexes from the Probabilistic Interaction Network

In order to define complex membership, we partitioned the high-confidence probabilistic physical interaction network by using the cluster growth algorithm ClusterONE (Nepusz et al., 2012), which outperformed other clustering methods on the denoised PPI network (Table S5). In total, the clustering predicts 622 discrete putative complexes encompassing 2,634 distinct proteins (Table S3). Complex membership size distribution approximated an inverse power law with a median of four subunits (Figure S4A). The majority (62%; 385/622) of the complexes have not been annotated (i.e., only 237 are currently curated in a public database like CORUM; Figures 3A and 3C). Although the fraction of curated components varies, we also recapitulated 258 previously reported complexes (Figure 3C), including several well-known membrane-associated complexes, such as the coat protein I and II (COPI/II) vesicle transport complexes that shuttle cargo between the Golgi and endoplasmic reticulum. Strikingly, most (67%; 335) of the 500 smaller putative complexes with five or fewer components, including the bulk (74%; 83) of the 112 predicted heterodimers, have never been curated before (Figure 3C).

Both independent experimental validation based on more traditional immunoprecipitation or coaffinity purification methods



D

	Mitocheck PPIs	Co-regulator Complexome	Fly Co-AP Interactions	Fly Co-AP Clusters
Annotated edges	281	13,663	8,688	8,688
Edges sharing annotation	131	2,873	1,549	536
Fold-enrichment over chance	4	477	9	87
Hypergeometric p-value	*1.69E-52*	*<E-308*	*<E-308*	*<E-308*

Figure 3. Global Validations of the Map of High-Confidence Human Protein Complexes

(A) Complex size distribution of the 622 inferred complexes.

(B) Network of predicted human protein complexes proportioned according to subunit number and displaying existing curations, validation status by AP/MS (Malovannaya et al., 2011), and PPI connectivity (proportioned edge width).

(C) Proportions of annotated complexes in public repositories (CORUM, PINdb, REACTOME, and HPRD) or independently experimentally verified.

(D) Enrichment analysis showing overlap with large-scale APMS data sets generated for human (Hutchins et al., 2010; Malovannaya et al., 2011) and (via orthology) fly (Guruharsha et al., 2011).

See also Table S3.

and orthology mapping support at least 21 of these putative complexes (i.e., not in any reference database) (Table S3; see Supplemental Information for details). For example, Guruharsha et al. (2011) recently reported 299 cocomplex interactions based on pull-down experiments of 43 affinity-tagged human proteins present in 41 of our complexes, of which 143 interactions map precisely to our predicted complexes, representing a 47.8% validation rate (which may be an underestimate, as Guruharsha et al. [2011] do not report human interactions that fall outside the fly interologs examined in their study). Likewise, the results of Malovannaya et al. (2011), who used large-scale immunoprecipitation to isolate native human protein complexes, show excellent agreement to 123 of our complexes (i.e., Benjamini-corrected hypergeometric $p \leq 0.05$), including 42 (34%) of our complexes that are not curated in CORUM (Figure 3B and Table S3). Figure 3D summarizes the highly significant overlap of our inferred complexes with these fully independent data sets, with enrichments ranging from 4- to 477-fold more than chance, thus broadly and systematically validating our network of derived human protein complexes.

By design, insoluble membrane-associated (hydrophobic) protein complexes were largely missed in this study, and the proteins assigned to complexes had a higher average transcript abundance (Figures S2A and S2B). Moreover, in an effort to control the false positive rate, our conservative clustering algorithm, ClusterONE, underweighted small clusters of size 2 or 3 for lack of sufficient association evidence, likely contributing to the prominence of complexes with four subunits in Figure 3A. But we did not observe any significant bias toward negative ($pI \leq 7$) or positive ($pI \geq 7$) charge as compared to complexes curated in CORUM (Figure S4B).

Figure 4 shows the broad functional diversity of the predicted complexes (a navigable map is available online for close visualization of individual clusters and their supporting cocomplex interactions). Consistent with biological expectation (Hartwell et al., 1999; Lage et al., 2007; Oliver, 2000; Vidal et al., 2011),

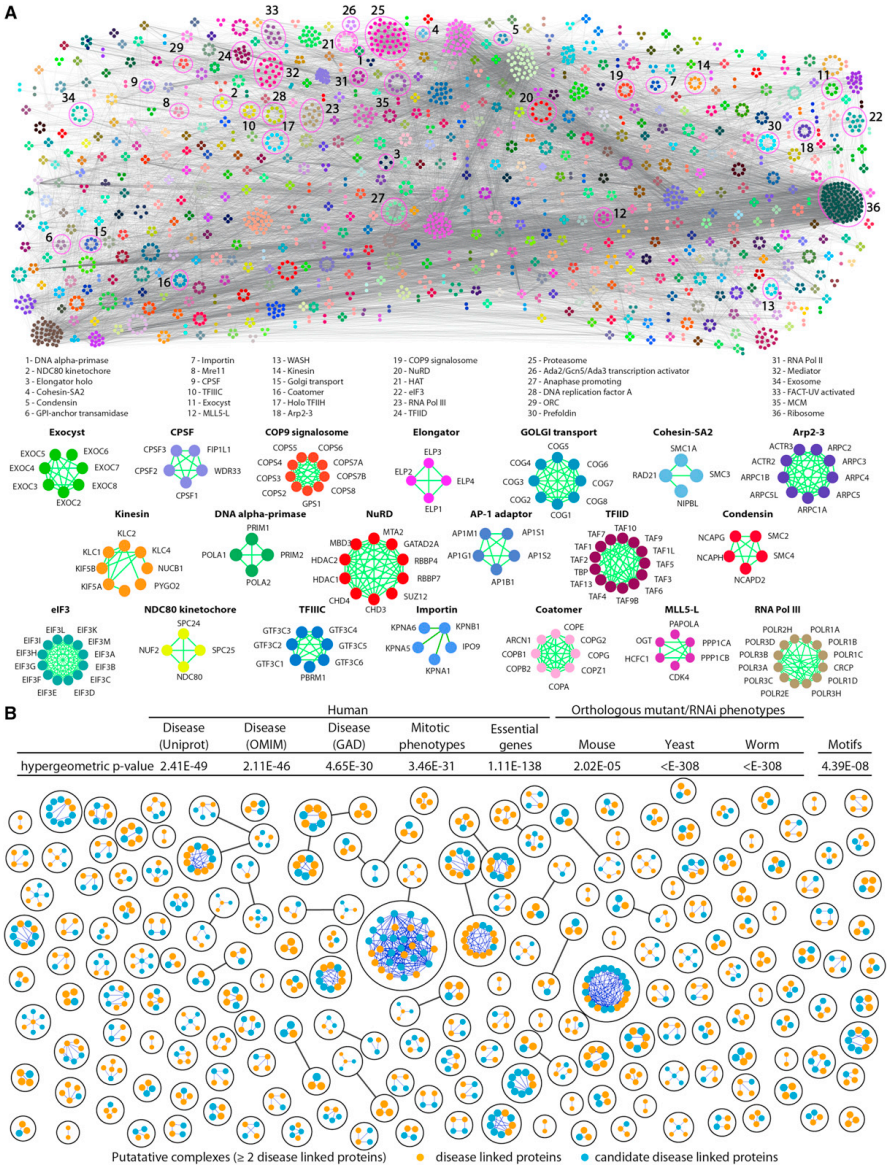


Figure 4. Global Map of High-Confidence Human Protein Complexes

(A) Schematic of the global network of inferred human soluble protein complexes (colored by membership) with representative examples and supporting PPI highlighted.

(B) Putative complexes with two or more components with human disorder associations annotated in UniProt (UniProt Consortium, 2011), Online Inheritance of Man (OMIM) (Hamosh et al., 2005), or the Genetic Association Database (GAD) (Becker et al., 2004). Inset table shows highly significant interaction overlap

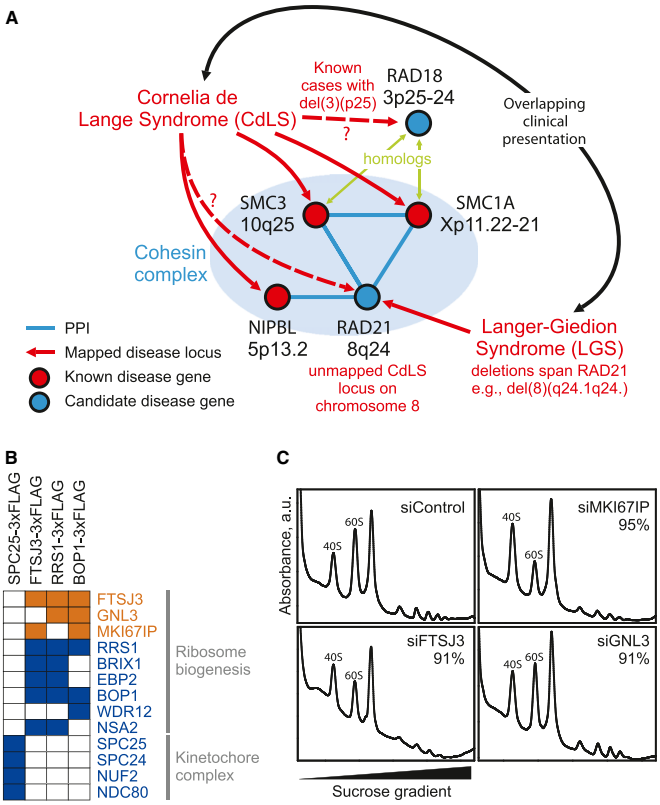


Figure 5. Membership in Complexes Predicts Protein Function and Disease Associations

(A) Three of four proteins mapped to the cohesin complex account for roughly half of cases of the human congenital disorder Cornelia de Lange syndrome (Pié et al., 2010), implicating the fourth component, RAD21, as a candidate disease gene. This association may explain similarities in clinical presentation between CdLS and Langer-Giedion syndrome, as the latter patients routinely harbor RAD21 deletions, e.g., McBrien et al. (2008) and Wuyts et al. (2002).

(B) Confirmation of ribosome biogenesis candidate (orange) associations with annotated components (blue) by AP/MS analysis of tagged proteins (top). Colored squares indicate validation (see Extended Experimental Procedures).

(C) Polysome profiling after siRNA targeting in tissue culture supports functional roles in ribosome biogenesis for three candidate proteins. Knockdown of MKI67IP, FTSJ3, and, to a lesser extent, GNL3, results in 60S ribosomal subunit biogenesis defects manifested by a reduced ratio of free 60S to 40S ribosomal subunits during gradient sedimentation as compared to control. Percentages indicate siRNA knockdown efficiency as measured by qRT-PCR.

the subunits of the complexes were significantly enriched for related biological functions, transcriptional regulatory motifs, and pathological processes (Figure 4B, inset table). Compared to the entire set of identified proteins, the clustered proteins also showed enrichment for posttranslation modifications linked to cellular regulation, like acetylation (Benjamini-corrected $p \leq 10^{-41}$) and phosphorylation ($p \leq 10^{-5}$). Many of the complexes are linked to core cellular processes, such as mRNA splicing ($p \leq 10^{-15}$) or transcription ($p \leq 10^{-5}$), that either are essential in human ($p \leq 10^{-138}$) or that have RNA interference (RNAi)-induced phenotype in cell culture (e.g., cell division arrest, $p \leq 10^{-31}$) or are associated, via orthology, with similar mouse, yeast, or worm mutant phenotypes (Figure 4B, inset table; see Table S4 for details).

Clinical and Biological Implications of the Reconstructed Human Protein Complexes

Consistent with this strong tendency for proteins in the same complex to be affiliated with similar mutational and RNAi phenotypes, subunits of the predicted human protein complexes were much more likely than chance ($p \leq 10^{-46}$) to have links to a documented clinical pathology (Figure 4B; see Table S4 for details), with disease-associated proteins distributed broadly among the complexes (Figures 4B and S4C). Closer examination of the interaction subnetworks—comprising known human disease genes with genes that currently lack annotation or that have not previously been associated with any human disorders (Figure 4B)—highlights the utility of the map.

One such example is shown in Figure 5A, illustrating the case of the human developmental disorder Cornelia de Lange syndrome (CdLS). Mutations in three subunits of the cohesin complex (SMC1A, SMC3, and NIPBL) have been linked to CdLS (Pié et al., 2010), implicating an additional component (RAD21) as a candidate CdLS locus, and are consistent with at least one unmapped CdLS locus residing on chromosome 8 (DeScipio et al., 2005). The link to RAD21 provides a likely

(i.e., shared annotated edges) with phenotypic data sets that reveals that protein subunits of the same predicted human complex tend to exhibit similar disease and genetic associations in human populations (see Extended Experimental Procedures), RNAi phenotypes in cell culture (Neumann et al., 2010), mutational and RNAi phenotypes in other species (via orthology), and shared transcriptional regulatory motifs (Xie et al., 2005).
See also Figure S4C and Table S4.

explanation for the occasional overlap of Langer-Giedion syndrome (LGS) clinical presentation with CdLS, as all LGS patients are at least partially defective for RAD21 (see e.g., McBrien et al., 2008; Wuyts et al., 2002). Similarly, RAD18, a homolog of SMC3 and SMC1A, may play a role in CdLS that is consistent with unmapped CdLS deletions within chromosome 3p25 (DeScipio et al., 2005). Reports coinciding with the preparation of this manuscript confirm that RAD21 mutations do indeed lead to a CdLS-like syndrome (Deardorff et al., 2012), supporting the use of the complex map to prioritize promising candidate genes for human diseases.

Similarly, participation in the same complex suggests shared functions; the map can thus be used to predict new biochemical functions for proteins and other types of functions. We experimentally validated one such case for a ribosome-associated subcomplex containing BOP1, RRS1, GNL3, EBP2, FTSJ3, and MK1671P, and we first confirmed the interactions by affinity tagging/purification and mass spectrometry (Figure 5B). BOP1, EBP2, and the yeast ortholog of RRS1 are known to participate in maturation of the large 60S ribosomal subunit, suggesting that the other factors likewise engage in ribosome assembly, which is consistent with the nucleolar localizations of GNL3, FTSJ3, and MKI67IP. Supporting a role in ribosome biogenesis, short interfering RNA knockdowns of FTSJ3, MKI67IP, and, to a lesser extent, GNL3, perturbed 60S formation in cell culture, decreasing the ratio of free 60S to 40S subunits (Figure 5C). Taken together, these data support roles in ribosome biogenesis for these proteins and confirm the utility of the map for identifying biological functions.

Conservation of Human Protein Complexes

Estimates based on sequence similarity across orthologs indicate that the components of the complexes we detect are generally more ancient and have higher conservation on average than most human proteins (Figure 6A; see Table S3 for details). Using orthology relationships derived from well-established sources and calculating evolutionary rates and ages for all human proteins as a base distribution for gauging the emergence of complexes (see Extended Experimental Procedures), we found that many complexes appear to be quite ancient and slowly evolving (Figure 6B). Strikingly, however, most (60%; 376/622) human complexes likely arose with vertebrates, i.e., orthologs not present in invertebrates or fungi (Table S3). Hence, our analyses suggest a major shift/expansion in the ancestral protein interaction network coincident with the emergence of vertebrates.

Given the availability of experimentally derived networks of fly and yeast protein complexes, we could directly examine the evolutionary conservation of protein complexes across animals by comparing our network of human complexes with the extensive maps of 556 fly protein complexes recently reported for *D. melanogaster* (Guruharsha et al., 2011) and 720 yeast protein complexes documented for *S. cerevisiae* (Babu et al., 2012). Roughly one quarter (24%; 149/622) of the predicted human protein complexes showed statistically significant overlaps with complexes reported for these models (Figure 6B, inset; see Table S3 for details), with half of the subunits having clear orthologs (Figure 6C); the remaining components presumably represent genuine differences or incomplete orthology annotations.

The functional significance of unannotated ancestral human complexes supported by conservation in yeast or fly (Table S3 and Figure 6) warrants further investigations. At least one such complex, a multisubunit transfer RNA (tRNA)-splicing ligase (Popow et al., 2011), was characterized recently. The interaction between DDX1 and C14orf166 was detected at high confidence both in our data set (probability score 0.899) and in the Guruharsha et al. (2011) fly cocomplex data, and the other respective associated complex subunits likewise show significant overlap (Benjamini-corrected p value 1.1×10^{-7}). Additional examples of complex conservation are similarly supported by independent experimental evidence, e.g., such as the matching tissue specificities of the putatively interacting proteins endoplasmin and glucosidase 2β (Figure 6D), which form an uncharacterized complex conserved in both the fly and human maps.

Functional enrichment analysis of ancient complexes in comparison to vertebrate-specific ones also reveals intriguing biological trends. For example, we expected ancient, core cellular functions to be depleted among vertebrate-specific complexes. Consistent with this expectation, we find proteins associated with the ribosome ($p \leq 10^{-67}$, 113 proteins) and RNA polymerase II ($p \leq 10^{-27}$, 45 proteins) to be highly enriched only among conserved complexes. However, we also observe several notable variations from this hypothesis. For example, compared to the genomic background, mitochondrial proteins are more highly enriched among proteins assigned to vertebrate complexes than among those assigned to conserved complexes; 159 vertebrate proteins have a mitochondrial Gene Ontology Biological Process (GO BP) annotation ($p \leq 10^{-31}$) versus only 81 proteins assigned to conserved complexes ($p \leq 10^{-5}$). Similarly, proteins annotated as being part of the splicing apparatus are enriched in both conserved ($p \leq 10^{-33}$; 63 proteins) and vertebrate complexes ($p \leq 10^{-11}$, 43 proteins), which is consistent with an ancient function gaining additional complexity in vertebrates (e.g., increased alternative splicing). Our study therefore offers a unique perspective into the functional conservation and diversification of protein complexes across animals.

Protein Abundance, Ubiquity, and Complex Subunit Stoichiometries

Consistent with the documented origins of the HeLa and HEK293 cells analyzed in this study, the complexes we identified were significantly enriched for epithelial markers ($p \leq 10^{-183}$; UniProt tissue annotations). Explicit comparison of results across the two cell lines used in this study provided little evidence for tissue-specific or cell-type-specific complexes (see Supplemental Information). Most proteins were detected in both cell line fractionations, which is consistent with the similar protein and mRNA expression patterns observed in these cell lines (Figure S1), whereas the few proteins detected uniquely in one cell line or the other did not preferentially assort into tissue-specific complexes (Figure S2). The vast majority of complex components are universally expressed in 11 cancer cell lines (Geiger et al., 2012) (Figure S3A) and show high and largely invariant expression in an mRNA sequencing (mRNA-seq) study of 16 normal human tissues (EBI accession number

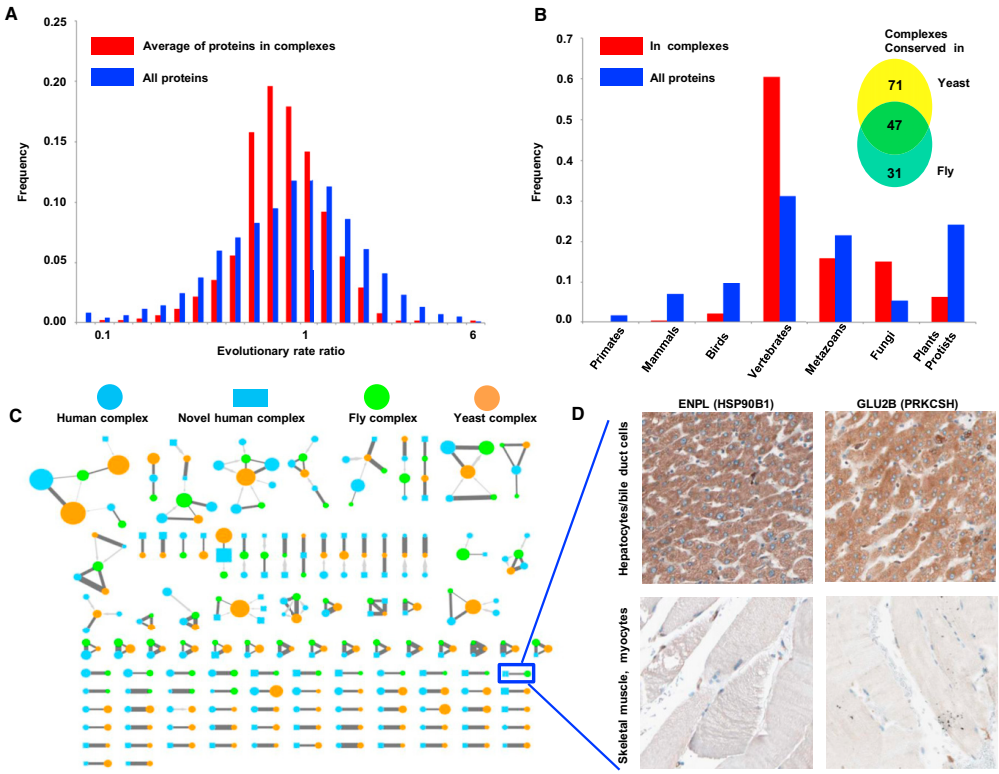


Figure 6. Evolutionary Conservation of Protein Complexes

(A) Components of predicted human complexes—calculated as the average of evolutionary rate ratios—evolved more slowly, as compared to the entire set of expressed proteins (see Extended Experimental Procedures).

(B) Pronounced spike in number of complexes originated with the emergence of vertebrates. x axis shows increasingly inclusive orthologous groups in the phylogeny of eukaryotes.

(C) Human complexes conserved in fly (Guruharsha et al., 2011) and yeast (Babu et al., 2012) (see Table S3 and Extended Experimental Procedures). Nodes represent complexes (human, blue; fly, green; yeast, orange), with size proportional to subunit number. Reciprocal best matches shown as dark gray edges, and nonreciprocal is shown as lighter gray directed edges, with edge thickness proportional to Sorensen-Dice overlap of complex members. Human complexes absent from public databases (putative complexes) are drawn as rectangles, and the remaining are drawn as circles.

(D) Similar tissue-specific expression patterns support a functional association between interacting proteins ENPL and GLU2B, whose orthologs were reported to interact in fly (Guruharsha et al., 2011). Panels show representative antibody staining in normal tissue biopsies collected and reported by the Human Protein Atlas (Uhlen et al., 2010) (www.proteinatlas.org).

See also Figure S3 and Table S3.

E-MTAB-513) (Figure S3B). Indeed, complex subunits are considered near ubiquitous ($p \leq 10^{-11}$; protein information resource [PIR] tissue specificity annotations) and are expressed in the top quartiles of 1,045 of 7,067 neoplastic and normal tissue CGAP EST libraries (1% false discovery rate [FDR]), including normal kidney ($p \leq 10^{-39}$), muscle ($p \leq 10^{-20}$), liver ($p \leq 10^{-12}$), brain ($p \leq 10^{-20}$), vascular ($p \leq 10^{-30}$), bone ($p \leq 10^{-15}$), and embryonic tissue ($p \leq 10^{-31}$). Consistent with this, genes encoding complex subunits also tend to share common upstream transcriptional regulatory motifs ($p \leq 10^{-8}$) (Figure 4B, inset table). Proteins mapped to complexes showed no major bias in abundance over the complete set of human proteins identified by mass spectrometry (Figure 1D).

The pervasiveness of ubiquitously expressed protein complexes argues strongly for broad relevance to basic human cell biology. Although often coexpressed, the subunit stoichiometries of human protein complexes in vivo are largely unknown and have never been systematically measured globally. Because

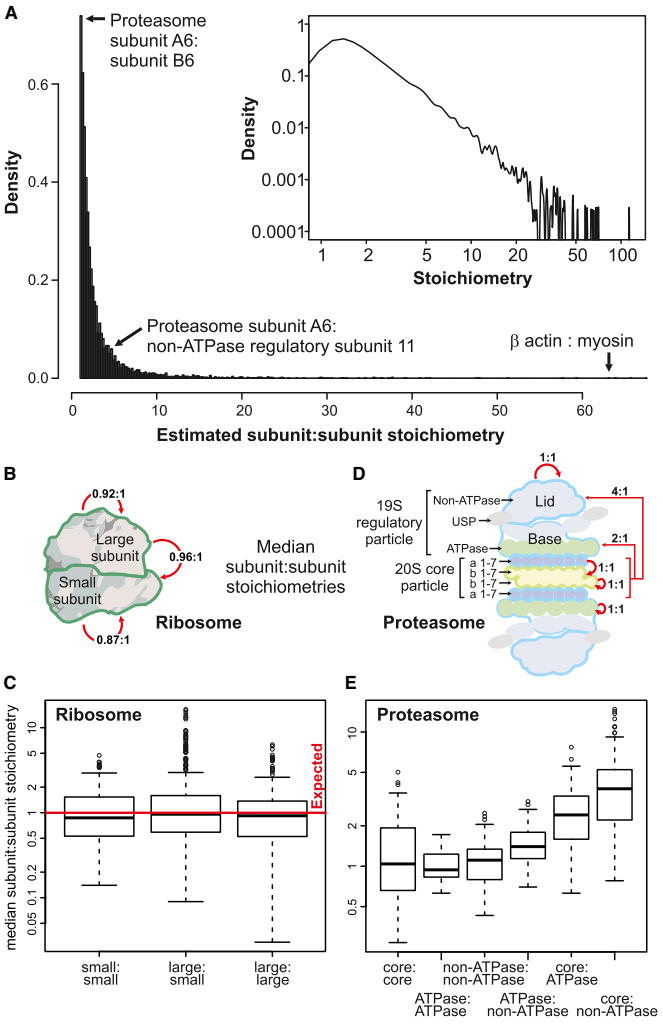


Figure 7. Protein Complex Stoichiometries

(A) Overall distribution of derived intracomplex component stoichiometries.

(B and C) Estimated subunit stoichiometries within and between proteins of the large and small ribosome subunits agree on average with the expected 1:1 ratio. Boxes summarize first quartile, median, and third quartiles, whiskers represent ±1.5 IQR, and circles represent outliers.

(D and E) Estimated protein subunit stoichiometries within and between proteasomal proteins. Intrasubunit stoichiometries within the core, ATPase, or nonATPase regulatory subunits agree well with the expected 1:1 ratio, but stoichiometries observed between these complexes deviate significantly from 1:1 (ATPase:non-ATPase, Mann-Whitney $p \leq 10^{-3}$; core:ATPase, $p \leq 10^{-12}$; core:non-ATPase, $p \leq 10^{-16}$).

See also Table S2.

all reconstructed complexes are supported by the same set of extensive experimental mass spectrometry data, we could estimate subunit stoichiometries based on the ratios of recorded spectral counts after correcting appropriately for protein size and composition (see Extended Experimental Procedures). Although only approximate ratios were inferred and peaked at ~1:1 (Figure 7A), such as between known ribosomal subunits (Figures 7B and 7C), the results highlight intriguing deviations in subunit abundance (Table S2). An example drawn from the proteasome is illustrative: whereas the median stoichiometry of core α and β enzymatic subunits is close to the expected 1:1 ratio, the median of stoichiometries of core to non-ATPase regulatory subunits deviated significantly at ~4:1 (Mann-Whitney $p \leq 10^{-16}$; Figures 7D and 7E). Hence, these data suggest a rich source of information about the physical organization of human proteins.

DISCUSSION

The biochemically based interaction data obtained in this integrative proteomic study have enabled the identification of both 364 previously unannotated protein complexes (i.e., predicted complexes with no statistically significant match to complexes in public databases) encompassing 1,278 human proteins, many of which are linked to human disease, and unexpected components and interactions for well-studied, widely conserved nuclear and cytoplasmic protein machineries, such as ribosome biogenesis, with clear biological implications. Most of the high-confidence protein interactions provided in this resource have not been previously reported in public interaction databases and hence motivate mechanistic investigations of specific biological systems.

Prior to this work, experimental knowledge regarding soluble protein complex membership in human cells has generally been ad hoc or focused on specific subcellular systems. Our relatively unbiased integrative approach, wherein biochemical evidence (cofractionation) of soluble native macromolecules was combined with genomic inferences (imputed functional associations), provides an inclusive snapshot of human protein

complexes present under a standardized cellular context, thus serving as a reference against which future process- or cell-type-specific or dynamic interaction data sets can be compared.

Information gleaned from orthology proved to be an important resource in separating true pairwise interactions from putative false positives and, in turn, could reasonably be expected to bias our results toward conserved complexes. In fact, although we do find conserved complexes as expected, we also find a majority that are not conserved (in fly and yeast) and that seemingly have arisen with vertebrates (i.e., Figure 6B). The slower rate of evolution of the subunits we report for our protein complexes is also a feature of other human PPI networks, such as in CORUM, and thus, our predictions of broad complex conservation, albeit incomplete, are not just artifacts of our methodology.

The fact that we detected little evidence of tissue specificity for most of the derived human protein complexes and few cell-type-specific components likely reflects undersampling by our mass spectrometry procedures, which is a common limitation of LC-MS/MS. At the level of predicted PPI (which are derived from multiple biochemical fractions), differences in the proteomic profiles generated for the two cell lines lie within the variance observed between biological replicates of the same cell line (Figures S1 and S2). Yet it is clear that differential interactomes and the contextual rewiring of PPI networks are major determinants of cell behavior and phenotypes. The complexes we report undoubtedly undergo differential rewiring in response to environmental, physiological, developmental, or disease states. With further refinements to our experimental procedures, our interaction mapping strategy has the potential to interrogate changes in interaction space in a systematic manner in the future.

To enable exploitation of these data by the scientific community, we have generated a dedicated web database of human protein complexes (http://human.med.utoronto.ca) that contains all the data generated in this study in an easily navigated format. These include all of the supporting information for each of the pairwise protein interactions obtained through integration of our cofractionation data with public genomic evidence, a list of the 5,584 proteins detected in each of the 1,163 biochemical fractions collected, and the subunit composition of the 622 putative protein complexes obtained through clustering of our generated high-confidence interaction network. This "first pass" draft of the soluble, stably associated human protein "complexome" provides a glimpse into the global physical molecular organization of human cells, which is likely to be perturbed in pathological states.

EXPERIMENTAL PROCEDURES

Cell Culture and Extract Preparation

HeLa S3 (ATCC#: CCL-2.2) and HEK293 (ATCC#: CRL-1573) soluble nuclear and cytoplasmic protein extracts were prepared by conventional methods (see Extended Experimental Procedures). Prior to fractionation, lysates were treated with 100 units/ml Benzonase (Novagen Inc.) to remove nucleic acids and clarified by centrifugation to remove debris.

Biochemical Fractionation and Proteomic Analysis

We performed weak anion-exchange and mixed-bed ion exchange, both with and without a heparin precolumn to enrich for nucleic-acid-binding proteins. In total, 1,095 chromatography fractions were collected (see Extended Experimental Procedures). Isoelectric focusing was carried out on a MicroRotofor Liquid-Phase IEF cell (Bio-Rad) according to the manufacturer's protocol, with 40 fractions collected across a pH range. Sucrose density gradient centrifugation was performed as previously described (Ramani et al., 2008), with 28 fractions collected.

Proteins were acid precipitated and trypsin digested, and the peptide mixtures were fractionated and sequenced by using nanoflow liquid chromatography-electrospray tandem mass spectrometry. Spectra were collected on an LTQ linear ion trap (ThermoFisher Scientific) (majority) or LTQ Orbitrap Velos hybrid mass spectrometer and searched against a UniProt human target-decoy sequence database by using SEQUEST (Eng et al., 2008) (see Extended Experimental Procedures). The LC-MS/MS identifications were filtered to a 1.0% protein and peptide theoretical FDR.

Bioinformatics Analyses

Protein cofractionation networks were scored by correlation analysis (Pearson correlation, weighted cross-correlation, coapex) based on the protein spectral counts recorded across each set of fractions (see Extended Experimental Procedures). Weighted networks were likewise constructed based on functional evidence reported in HumanNet (Lee et al., 2011), omitting human protein interaction data to minimize circularity that might bias our association predictions. A coevolution network (Tillier and Charlebois, 2009) based on correlated evolutionary rates was built to account for additional associations not covered in HumanNet.

For the machine-learning classifier, we used the fast random forest implementation in Weka (see Extended Experimental Procedures) to integrate all generated networks. Cross-validated decision trees were learned and benchmarked by using independent training and test sets of CORUM reference complexes (Ruepp et al., 2010). We denoised the network by using a diffusion procedure to delete interactions lacking network topology support and by calibrating the diffused interaction scores with Gene Ontology (Cellular Component) normalized semantic similarity scores (see Extended Experimental Procedures).

Clusters were defined by using the ClusterONE algorithm with parameter settings chosen to yield the highest maximum matching ratio (Nepusz et al., 2012) between the predicted complexes and set of cluster-training complexes (see Extended Experimental Procedures).

Stoichiometries calculation is shown in Extended Experimental Procedures.

ACCESSION NUMBERS

The interaction data have been deposited into BioGRID and are also publicly accessible via a dedicated web portal (http://human.med.utoronto.ca/).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, five figures, and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2012.08.011.

ACKNOWLEDGMENTS

We thank R. Isserlin, Z. Ni, H. Guo, D. Merico and A. Alpert for technical assistance and J. Parkinson, G. Bader, A. Wilde, and J. Greenblatt for critical suggestions. P.C.H. was a recipient of a University of Toronto Open Fellowship, T.N. was supported by the Newton International Fellowship Scheme of the Royal Society, A.E. is an Ontario Research Chair, and S.J.W. is a Canada Research Chair Tier 1. This work was supported by grants from the Biotechnology and Biological Sciences Research Council (BB/F00964X/1 and BB/K004131/1) and the Royal Society (NF080750) to A.P., the Canada Institutes of Health Research (MOP 82940) and the SickKids Foundation to S.J.W., the National Institutes of Health, National Science Foundation, Cancer Prevention Research Institute of Texas, and Welch (F1515) and Packard Foundations to E.M.M., and the Ontario Ministry of Research and Innovation to A.E.

Received: May 26, 2012
Revised: July 30, 2012
Accepted: August 10, 2012
Published: August 30, 2012

REFERENCES

Alberts, B. (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell *92*, 291–294.

Babu, M., Vlasblom, J., Pu, S., Guo, X., Graham, C., Bean, B.D.M., Vizeacoumar, F.J., Burston, H.E., Snider, J., Phanse, S., et al. (2012). Interaction Landscape of Membrane Protein Complexes in Saccharomyces cerevisiae. Nature http://dx.doi.org/10.1038/nature11354.

Becker, K.G., Barnes, K.C., Bright, T.J., and Wang, S.A. (2004). The genetic association database. Nat. Genet. *36*, 431–432.

Behrends, C., Sowa, M.E., Gygi, S.P., and Harper, J.W. (2010). Network organization of the human autophagy system. Nature *466*, 68–76.

Bouwmeester, T., Bauch, A., Ruffner, H., Angrand, P.O., Bergamini, G., Croughton, K., Cruciat, C., Eberhard, D., Gagneur, J., Ghidelli, S., et al. (2004). A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway. Nat. Cell Biol. *6*, 97–105.

Butland, G., Peregrín-Alvarez, J.M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., et al. (2005). Interaction network containing conserved and essential protein complexes in Escherichia coli. Nature *433*, 531–537.

Deardorff, M.A., Wilde, J.J., Albrecht, M., Dickinson, E., Tennstedt, S., Braunholz, D., Mönnich, M., Yan, Y., Xu, W., Gil-Rodríguez, M.C., et al. (2012). RAD21 mutations cause a human cohesinopathy. Am. J. Hum. Genet. *90*, 1014–1027.

DeScipio, C., Kaur, M., Yaeger, D., Innis, J.W., Spinner, N.B., Jackson, L.G., and Krantz, I.D. (2005). Chromosome rearrangements in cornelia de Lange syndrome (CdLS): report of a der(3)t(3;12)(p25.3;p13.3) in two half sibs with features of CdLS and review of reported CdLS cases with chromosome rearrangements. Am. J. Med. Genet. A. *137A*, 276–282.

Eng, J.K., Fischer, B., Grossmann, J., and Maccoss, M.J. (2008). A fast SEQUEST cross correlation algorithm. J. Proteome Res. *7*, 4598–4602.

Ewing, R.M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M.D., O'Connor, L., Li, M., et al. (2007). Large-scale mapping of human protein-protein interactions by mass spectrometry. Mol. Syst. Biol. *3*, 89.

Gavin, A.C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature *415*, 141–147.

Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dümpelfeld, B., et al. (2006). Proteome survey reveals modularity of the yeast cell machinery. Nature *440*, 631–636.

Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. (2012). Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Mol. Cell. Proteomics *11*, M111, 014050.

Graham, F.L., Smiley, J., Russell, W.C., and Nairn, R. (1977). Characteristics of a human cell line transformed by DNA from human adenovirus type 5. J. Gen. Virol. *36*, 59–74.

Guruharsha, K.G., Rual, J.F., Zhai, B., Mintseris, J., Vaidya, P., Vaidya, N., Beekman, C., Wong, C., Rhee, D.Y., Cenaj, O., et al. (2011). A protein complex network of Drosophila melanogaster. Cell *147*, 690–703.

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. *33*(Database issue), D514–D517.

Hartwell, L.H., Hopfield, J.J., Leibler, S., and Murray, A.W. (1999). From molecular to modular cell biology. Nature *402*(6761, Suppl), C47–C52.

Havugimana, P.C., Wong, P., and Emili, A. (2007). Improved proteomic discovery by sample pre-fractionation using dual-column ion-exchange high performance liquid chromatography. J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. *847*, 54–61.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., et al. (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature *415*, 180–183.

Hu, P., Janga, S.C., Babu, M., Díaz-Mejía, J.J., Butland, G., Yang, W., Pogoutse, O., Guo, X., Phanse, S., Wong, P., et al. (2009). Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. PLoS Biol. *7*, e96.

Hutchins, J.R., Toyoda, Y., Hegemann, B., Poser, I., Hériché, J.K., Sykora, M.M., Augsburg, M., Hudecz, O., Buschhorn, B.A., Bulkescher, J., et al. (2010). Systematic analysis of human protein complexes identifies chromosome segregation proteins. Science *328*, 593–599.

Jansen, R., and Gerstein, M. (2004). Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. Curr. Opin. Microbiol. *7*, 535–545.

Jeronimo, C., Forget, D., Bouchard, A., Li, Q., Chua, G., Poitras, C., Thérien, C., Bergeron, D., Bourassa, S., Greenblatt, J., et al. (2007). Systematic analysis of the protein interaction network for the human transcription machinery reveals the identity of the 7SK capping enzyme. Mol. Cell *27*, 262–274.

Kislinger, T., Rahman, K., Radulovic, D., Cox, B., Rossant, J., and Emili, A. (2003). PRISM, a generic large scale proteomic investigation strategy for mammals. Mol. Cell. Proteomics *2*, 96–106.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., et al. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature *440*, 637–643.

Kühner, S., van Noort, V., Betts, M.J., Leo-Macias, A., Batisse, C., Rode, M., Yamada, T., Maier, T., Bader, S., Beltran-Alvarez, P., et al. (2009). Proteome organization in a genome-reduced bacterium. Science *326*, 1235–1240.

Lage, K., Karlberg, E.O., Størling, Z.M., Olason, P.I., Pedersen, A.G., Rigina, O., Hinsby, A.M., Tümer, Z., Pociot, F., Tommerup, N., et al. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat. Biotechnol. *25*, 309–316.

Lee, I., Blom, U.M., Wang, P.I., Shim, J.E., and Marcotte, E.M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. *21*, 1109–1121.

Mak, A.B., Ni, Z., Hewel, J.A., Chen, G.I., Zhong, G., Karamboulas, K., Blakely, K., Smiley, S., Marcon, E., Roudeva, D., et al. (2010). A lentiviral functional proteomics approach identifies chromatin remodeling complexes important for the induction of pluripotency. Mol. Cell. Proteomics *9*, 811–823.

Malovannaya, A., Lanz, R.B., Jung, S.Y., Bulynko, Y., Le, N.T., Chan, D.W., Ding, C., Shi, Y., Yucer, N., Krenciute, G., et al. (2011). Analysis of the human endogenous coregulator complexome. Cell *145*, 787–799.

Masters, J.R. (2002). HeLa cells 50 years on: the good, the bad and the ugly. Nat. Rev. Cancer *2*, 315–319.

McBrien, J., Crolla, J.A., Huang, S., Kelleher, J., Gleeson, J., and Lynch, S.A. (2008). Further case of microdeletion of 8q24 with phenotype overlapping Langer-Giedion without TRPS1 deletion. Am. J. Med. Genet. A. *146A*, 1587–1592.

Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S., and Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. Mol. Syst. Biol. *7*, 548.

Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. Nat. Methods *9*, 471–472.

Neumann, B., Walter, T., Hériché, J.K., Bulkescher, J., Erfle, H., Conrad, C., Rogers, P., Poser, I., Held, M., Liebel, U., et al. (2010). Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. Nature *464*, 721–727.

Oliver, S. (2000). Guilt-by-association goes global. Nature *403*, 601–603.

Pesquita, C., Faria, D., Falcão, A.O., Lord, P., and Couto, F.M. (2009). Semantic similarity in biomedical ontologies. PLoS Comput. Biol. *5*, e1000443.

Pié, J., Gil-Rodríguez, M.C., Ciero, M., López-Viñas, E., Ribate, M.P., Arnedo, M., Deardorff, M.A., Puisac, B., Legarreta, J., de Karam, J.C., et al. (2010). Mutations and variants in the cohesion factor genes NIPBL, SMC1A, and SMC3 in a cohort of 30 unrelated patients with Cornelia de Lange syndrome. Am. J. Med. Genet. A. *152A*, 924–929.

Popow, J., Englert, M., Weitzer, S., Schleiffer, A., Mierzwa, B., Mechtler, K., Trowitzsch, S., Will, C.L., Lührmann, R., Söll, D., and Martinez, J. (2011). HSPC117 is the essential subunit of a human tRNA splicing ligase complex. Science *331*, 760–764.

Ramani, A.K., Li, Z., Hart, G.T., Carlson, M.W., Boutz, D.R., and Marcotte, E.M. (2008). A map of human protein interactions derived from co-expression of human mRNAs and their orthologs. Mol. Syst. Biol. *4*, 180.

Rhodes, D.R., Tomlins, S.A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A., and Chinnaiyan, A.M. (2005). Probabilistic model of the human protein-protein interaction network. Nat. Biotechnol. *23*, 951–959.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.W. (2010). CORUM: the comprehensive resource of mammalian protein complexes—2009. Nucleic Acids Res. *38*(Database issue), D497–D501.

Sardiu, M.E., Cai, Y., Jin, J., Swanson, S.K., Conaway, R.C., Conaway, J.W., Florens, L., and Washburn, M.P. (2008). Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. Proc. Natl. Acad. Sci. USA *105*, 1454–1459.

Sowa, M.E., Bennett, E.J., Gygi, S.P., and Harper, J.W. (2009). Defining the human deubiquitinating enzyme interaction landscape. Cell *138*, 389–403.

Tillier, E.R., and Charlebois, R.L. (2009). The human protein coevolution network. Genome Res. *19*, 1861–1871.

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., et al. (2010). Towards a knowledge-based Human Protein Atlas. Nat. Biotechnol. *28*, 1248–1250.

UniProt Consortium. (2011). Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res. *39*(Database issue), D214–D219.

Vidal, M., Cusick, M.E., and Barabási, A.L. (2011). Interactome networks and human disease. Cell *144*, 986–998.

Wessels, H.J., Vogel, R.O., van den Heuvel, L., Smeitink, J.A., Rodenburg, R.J., Nijtmans, L.G., and Farhoud, M.H. (2009). LC-MS/MS as an alternative for SDS-PAGE in blue native analysis of protein complexes. Proteomics *9*, 4221–4228.

Wuyts, W., Roland, D., Lüdecke, H.J., Wauters, J., Foulon, M., Van Hul, W., and Van Maldergem, L. (2002). Multiple exostoses, mental retardation, hypertrichosis, and brain abnormalities in a boy with a de novo 8q24 submicroscopic interstitial deletion. Am. J. Med. Genet. *113*, 326–332.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature *434*, 338–345.

Yang, H., Nepusz, T., and Paccanaro, A. (2012). Improving GO semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. Bioinformatics *28*, 1383–1389.

EXTENDED EXPERIMENTAL PROCEDURES

Biochemical Fractionation Using Native Chromatography

HPLC Columns, Buffers, and Instrumentation

IEX chromatography columns (weak anion-exchange PolyWAX LP; weak cation-exchange PolyCAT A; mixed-bed PolyCATWAX50/50 columns) were purchased from PolyLC Inc (MD, USA). TSKgel Heparin-5PW affinity column was obtained from Tosoh Bioscience LLC (PA, USA). Our buffer systems (fresh prepared with HPLC grade H_2O) comprised low salt buffer A [10 mM Tris-HCl, pH7.6, 3 mM NaN_3, 0.5 mM DTT, 5%-Glycerol] and high salt Buffer B [Buffer A + 1.5 M NaCl]. We performed all HPLC fractionations using an Agilent 1100 HPLC binary pump system (Agilent Technologies, ON, Canada), essentially described elsewhere (Havugimana et al., 2007). Protein elution was monitored by absorption at 280 nm.

Single-Phase Heparin Fractionation of Nuclear Extract

HeLa nuclear extract (~6.0 mg total proteins) prepared using traditional methods (Dignam et al., 1983) was fractionated on a TSKgel Heparin-5PW affinity column (75 × 7.5 mm id, 10 μm, 1000-A) previously equilibrated with buffer A at a flow rate of 0.5 ml/min. After loading, the bound proteins were eluted from the column with a 50 min gradient from 0 to 50% buffer B (buffer A + 1.5 M NaCl). A 5 min gradient with 50%–100% buffer B was applied to elute tightly bound proteins, with 100% buffer B maintained for an additional 3 min before returning back to 0% B for 7 min to re-equilibrate the column. In total, 48 × 0.75-ml fractions were collected from 0 to 72 min (1.5 min/fraction). Protein was precipitated with 10% TCA overnight at 4°C. The pellet was washed twice with −20°C acetone for 30 min. After air drying, the pellet was dissolved in 50 μl digest solution (50 mM NH_4HCO_3- 50 mM Tris, 1 mM $CaCl_2$). The sample reduction (room temperature, 1 hr) and alkylation (room temperature, 30 min) were respectively performed using 5 mM and 15 mM of Dithiothreitol and Iodoacetamide. Each protein fraction was digested with 1 μg of sequencing grade trypsin (Roche, Mississauga, Canada). After incubation for 18 hr at 30°C with gentle shaking (VWR incubating micro-plate shaker; 300 rpm) samples were dry speed-vac. 20 μl of LC-MS grade buffer (5% Formic Acid in HPLC grade water) were used to solubilise the peptide- digests. 8 μl tryptic peptides aliquot were directly analyzed by LC-MS.

Single-Phase Weak Anion-Exchange Fractionation of HeLa Cytosolic Extract

A total of 2.0-3.0 mg soluble protein in HeLa S3 cytosolic extract were applied to a PolyWAX LP column (200 × 4.6 mm id, 5 μm, 1000-A) equilibrated with buffer A. Elution of bound proteins was achieved through application of a 30 min gradient from 0 to 50% buffer B, with a final 2 min gradient of 50%–100% buffer B applied to elute tightly bound proteins. 100% buffer B was maintained for an additional 2 min before returning back to 0% buffer B in 2 min for re-equilibration of the column for 3 min. A total of 45 × 1.2-ml fractions were collected using a flow rate of 1.2 ml/min. The first and last fractions lacking protein (as judged by UV-absorption at 280 nm) were discarded. The rest of collected fractions were processed as described above.

Dual-Phase Heparin-Mixed-Bed Ion Exchange Fractionation of Nuclear Extracts

To enhance detection of low abundance nuclear proteins by MS, we used an optimized high resolution tandem affinity column coupled online with a mixed-bed ion exchange column to enrich and resolve multi-proteins complexes in nuclear extracts. Typically, 8-10 mg proteins from HeLa or HEK293 nuclear extracts were loaded on a dual TSKgel Heparin-5PW affinity column (75 × 7.5 mm id, 10 μm, 1000-A) coupled in series with PolyCATWAX mixed-bed ion exchange column (200 × 4.6 mm id, 12 μm, 1500-A) mounted to our integrated Agilent 1100 HPLC system (Agilent Technologies, ON, Canada). A 4 hr salt gradient (0.15 - 1.5 M NaCl) in Binding Buffer A was used at 0.25 ml/min to resolve and fractionate proteins into 120 × 0.5-ml time-based fractions for downstream MS protein identification. HeLa nuclear extract was fractionated in duplicates to confirm the reproducibility.

Triple-Phase Ion-Exchange Fractionation of HeLa Nuclear Extracts

As we have shown in our previous work (Havugimana et al., 2006, 2007), tandem weak anion-exchange (WAX) coupled in series to a weak cation-exchange (WCX) offered greater resolution than a single column or WCX-WAX in tandem. To minimize both chance co-elution and bias toward one chromatographic fractionation approach, we used our further semi-preparative optimized and reproducible triple phase IEX-HPLC that comprised our previously optimized columns system preceded with a long weak anion-exchange (250 × 9.4 mm i.d, 12 μm, 1500-A PolyWAX LP → 250 × 9.4-mm i.d, 12 μm, 1500-A PolyWAX LP → 250 × 9.4 mm i.d, 5 μm, 1500-A PolyCAT A) to fractionate 10-12 mg total proteins in HeLa nuclear extracts into 375 × 0.8-ml fractions using elution program consisting of a 10 min gradient with 100% buffer A to allow protein binding followed by a 50 min gradient with 0 to 50% buffer B followed by a10- min gradient with 50 to 100% buffer B, 10 min at 100% buffer B, 10 min with 100 to 0% buffer B, and finally 10- min at 100% buffer A to re-equilibrate the column for the next injection. A flow rate of 4-ml/min was applied in elution gradient program. Collected fractions were analyzed by LC-MS/MS in duplicates.

Triple-Phase Ion-Exchange Fractionation of HeLa Cytosolic Extracts

To identify macromolecular complexes that populate the HeLa cytoplasmic compartment, we scaled down our optimized semi-preparative IEX-HPLC fractionation procedure to enhance protein concentration in each collected fraction. Seven to 9 mg total proteins in HeLa cytoplasmic extract were fractionated on a triple phase IEX-HPLC analytical columns set up (200 × 4.6 mm i.d, 5 μm, 1000-A PolyWAX LP → 200 × 4.6-mm i.d, 5 μm, 1000-A PolyWAX LP → 200 × 4.6 mm i.d, 5 μm, 1000-A PolyCAT A) and resolved into 300 × 0.4-ml fractions using a 2.5 hr gradient elution program (23 min with 100% buffer A; 75 min with 0%–50% buffer B; 3 min with 50%–100% buffer B; 23 min with 100% buffer B; 3 min with 100 to 0% buffer B; 23 min with 100% buffer A) at flow rate of 0.8 ml/min. Both the 19 fractions representing the column flow through and the 12 fractions representing the re-equilibration step

were discarded as no proteins were detected in our short quality control LC-MS/MS analysis. All remaining 269 fractions were analyzed in duplicate by LC-MS/MS.

Biochemical Fractionation Using IEF and Sucrose Gradient Sedimentation

Sample Preparation for Isoelectric Focusing Fractionation

HeLa cells were grown to 70%–80% confluency in 75cm^2 flasks and harvested by mechanical scraping. Cells were washed in ice-cold PBS, pelleted by centrifugation (600xg), and resuspended in lysis buffer [10 mM Tris-HCl (pH 8.0), 10 mM KCl, 1.5 mM $MgCl_2$, 0.5 mM DTT, and 1x Protease Inhibitor Cocktail Set I (Calbiochem)]. Cells were lysed on ice using a Dounce homogenizer and fractionated into cytosolic and nuclear fractions using a protocol adapted from previous publication (Andersen et al., 2002). Briefly, cells were centrifuged at 1000xg for 5 min (4°C). The supernatant was saved as the cytosolic fraction. The pellet was resuspended in 250 mM sucrose/10 mM $MgCl_2$/1x Protease Inhibitor Cocktail, layered over a sucrose cushion of 880 mM sucrose/0.5 mM $MgCl_2$/1x Protease Inhibitor Cocktail, and centrifuged at 3000xg for 10 min (4°C). The supernatant was discarded and the pellet resuspended in lysis buffer with 5% NP-40 by sonicating water bath (15 min). Following sonication, samples were centrifuged at 3,500xg for 10 min to pellet insoluble material, with the supernatant saved as the nuclear fraction.

IEF Fractionation

Cytosolic and nuclear fractions were further fractionated in solution by isoelectric focusing on a MicroRotofor Liquid-Phase IEF cell (Bio-Rad). Ten fractions per sample were collected across a pH range of either 3-10 or 5-8. Following IEF fractionation, ampholytes were removed by OrgoSol DetergentOUT detergent removal kit (G-Biosciences).

Trypsin Digestion and MS Analysis of IEF Samples

Samples were denatured and reduced in 50% 2,2,2-trifluoroethanol (TFE) and 15 mM DTT at 55°C for 45 min, followed by alkylation with 55 mM iodoacetamide for 30 min at room temperature in the dark. Following alkylation, samples were diluted to 5% TFE in 50 mM Tris-HCl, pH8.0/2 mM $CaCl_2$ and digested with a 1:50 final concentration of Proteomics Grade trypsin (Sigma) for 5 hr at 37°C. Digestion was quenched by addition of 1% formic acid, and the sample volume was reduced to near dry (<20 μl) by speed vac centrifugation. Samples were resuspended in 5% acetonitrile/0.1% formic acid and bound and washed on HyperSep C18 SpinTips (Thermo). Following elution, the sample volume was reduced by speed vac to remove elution buffer. Samples were resuspended in 5% acetonitrile/0.1% formic acid and filtered through Amicon Ultra 10kDa centrifugation filters (Millipore).

Samples were analyzed by LC-MS/MS. Peptides were separated on a Zorbax 300SB-C18 reverse phase column (0.075 × 150 mm, 3.5 μm; Agilent) with an elution gradient of 5%–38% acetonitrile over 230 min followed by 38%–100% over 15 min. Peptides were analyzed by nanoelectrospray ionization onto an LTQ Orbitrap mass spectrometer (Thermo Scientific). Parent mass scans (MS1) were collected at high resolution (100,000) with data dependent ion selection activated for ions of greater than +1 charge. Up to 12 ions per MS1 were selected for CID fragmentation spectrum acquisition (MS2), with ions selected twice within 30 s placed on a dynamic exclusion list for 45 s.

Sucrose Gradient Fractionation of HeLa

Generation of the sucrose density gradient fractions and MS analysis were described elsewhere (Andersen et al., 2002; Ramani et al., 2008). Briefly, they were generated using a 7%–47% continuous sucrose gradient and ultra-high-speed centrifugation of the supernatants from HeLa S3 cell-free extracts. Gradient fractions were analyzed by Mass Spectrometry with LTQ-Orbitrap hybrid mass spectrometer (ThermoFisher), and tandem mass spectra were searched as described below.

LC-MS/MS Separation and Identification of Chromatographic Peptide Fractions

For LC-MS/MS analysis of HPLC protein fractions, samples were overnight 10%-TCA precipitated and neat-cold acetone was used to wash the precipitates. Proteins were then resuspended in 50 μl of trypsin digestion buffer [50 mM Ammonium Bicarbonate, 1 mM $CaCl_2$, 50 mM Tris; pH7.8], subjected to reduction (10 mM DTT, 30 min, 30°C), alkylation (15 mM IAM, 60 min, 30°C in the dark), and digestion (18 hr, 30°C, with gentle agitating) with one μg trypsin sequencing grade (Roche, Mississauga, Canada). The digestion mixture was dried in the Savant Speed Vacuum, and tryptic peptides were re-solubilised in 20 μl of 5% formic acid prior to analysis by LC-MS/MS using a linear ion trap mass spectrometer (LTQ; Thermo Fisher Scientific, CA, USA) or LTQ Orbitrap Velos (Thermo Fisher) coupled online to a nanoflow HPLC System (EASY-nLC; Proxeon, Odense, Denmark) via a nanoelectrospray ion source. Reverse-phase LC-MS/MS using 150-μm i.d × 40 cm in-house packed fused-silica C18 micro-capillary columns (Zorbax XDB-C18, 3.5 μm, Agilent Technologies, Canada) at a flow rate of 500 nl/min were used to resolve peptides mixture in each HPLC fraction. To separate peptides, we used columns with varying between 10-40 cm in length depending on sample complexity in each fractionation experiment. The gradient elution time was adjusted to the length of the column and varied between 2 and 4 hr. For a 2 hr gradient elution, 5 μl of tryptic peptides generated for TCS-HPLC fractions were loaded onto a 20-cm column and eluted with a 0 to 35% solvent B (0.1% formic acid/95% acetonitrile) over 90 min and from 35 to 95% in 15 min. For peptides analyzed on an LTQ ion trap instrument, eluted peptides were directly sprayed into an LTQ ion trap MS instrument via application of a spray voltage of 3.0 kV to a nanospray ion source (Proxeon). The MS was operated in a fully automated data-dependent manner using Xcaliber 2.0 software to acquire one full MS scan (400 - 2,000 m/z) followed by five MS/MS scans selected based on the five most abundant precursor ions and a precursor signal threshold of 1,000 counts. Ion fragmentation was performed in CID mode through application of normalized collision energy of 35%. Ions subjected to MS/MS were excluded from further sequencing for 30 s. For peptide mixtures analyzed on an LTQ-Orbitrap

Velos instrument, peptide samples were directly autosampled onto a 10 cm in-house packed column (75 μm inner diameter) with 3 μm reversed phase beads (Zorbx 80XDB-C18, Agilent). Using a 60 min gradient (5%–35% ACN), peptides were directly electro-sprayed (2.5 kV) into the mass spectrometer. Mass spectrometer was operated in data dependent mode switching automatically between one full scan MS and 10 MS/MS acquisitions. Instrument control was through Tune 2.6.0. and Xcalibur 2.1.0. Full scanMS spectra (400 – 2,000 m/z) were acquired in the Orbitrap analyzer after accumulation to a target value of 10^6 in the linear ion trap (resolution of 60,000 at 400 m/z).Fragmentation was performed in CID mode applying 35% normalized collision energy.

LC-MS/MS Spectra Database Search and Protein Identification

All MS/MS spectra IEF, SGF and IEX experiments acquired during over 9,000 hr of dedicated instrument run time were combined (resulting in > 18,000,000 mass spectra) and rigorously searched against a target-decoy human database downloaded from Universal Protein Resources Database (UniProtKB/Swiss-Prot Release 57.11; comprising 20,328 human proteins supplemented with common contaminants) using the SEQUEST algorithm (V2.7) as previously described (Eng et al., 2008). Static modifications were permitted to allow for the detection of carboxyamidomethylated (+57amu) cysteine. All peptide matches were required to be fully tryptic although one missed cleavage was permitted. The probabilistic STATQUEST model (Kislinger et al., 2003) was used to evaluate and assign confidence scores to all putative matches. Both proteins and peptides were considered positively identified if detected within a 1% false discovery rate cut off (based on empirical target-decoy database search results). The proteomic patterns of the HPLC, IEF and SGF fractions were compared using the CONTRAST software tool (Tabb et al., 2002). We then removed from consideration all proteins that passed our stringent cut off with only a single spectral count across all combined MS runs. Moreover, to ensure a high quality proteomic data set, we confirmed the expression of our LC-MS detected proteins by cross-comparing with previously reported HeLa S3 and HEK293 mRNA deep-sequencing data sets (Morin et al., 2008; Sultan et al., 2008). Additionally, we only kept proteins that were supported by at least two unique peptides in at least one recent comprehensive proteomic study of the HeLa proteome (Selbach et al., 2008; Wiśniewski et al., 2009). This screening procedure resulted in 41,506 unique peptides (supported by ~1.6 million individual mass spectra) matching to 5,584 distinct human proteins. To facilitate cross-mapping between data sets, we used UniProtKB accession numbers as a common identifier and the UniProt ID mapping tool to interconvert different gene and protein identifiers.

Polysome Profiling and Quantitative RT-PCR

HeLa cells were maintained in Dulbecco's Modified Eagle's Media (DMEM) supplemented with 10% fetal calf serum in a humidified 5% CO_2 incubator at 37°C. Cells were transfected with 10 nM ON-TARGETplus SMARTpool siRNA (Thermo Scientific Dharmacon) by using RNAiMAX (Invitrogen) at about 30% confluency. After 48 hr, 100 μg/ml cycloheximide (Sigma) was added into the culture medium and cells were incubated for 5 min in the incubator. Then cells were collected by trypsinization and washed with cold PBS containing 100 μg/ml cycloheximide twice. $1x10^5$ cells were frozen in −80°C for RNA extraction. The remaining cells were lysed in the lysis buffer (20 mM Tris, pH7.4, 100 mM KCl, 10 mM MgCl2, 1% Triton-100, 1 mM DTT, 100 μg/ml cycloheximide, 1x EDTA-free inhibitor tablet) on ice for 5 min. Extracts were clarified by centrifugation at 13,000 rpm for 10 min at 4 deg. The supernatant was loaded onto a linear sucrose gradient (15%–45%) prepared in lysis buffer without Triton. After a 4 hr centrifugation at 36,000 rpm in a Beckman SW40 rotor, the sucrose gradient was fractionated and absorbance at 254 nm was measured (ISCO fractionator). For qRT-PCR, total RNA was extracted by RNeasy Plus Micro (QIAGEN). QuantiTect reverse transcription kit and QuantiFast SYBR Green RT-PCR Kit from QIAGEN were used for qRT-PCR. The primer pairs for each gene in qRT-PCR were as follow: human MKI67IP(rMKI67IP-1: 5′-CCTGTTTGGTGAAAGACTCTTG-3′; rMKI67IP-2: 5′-GCTTTTGTGTTAGTGTCCGATTC-3′), Human GNL3(rGNL3-1: 5′-CATTCGGGTTGGAGTAATTGG-3′; rGNL3-2: 5′-TGTGATCTGTTTGTCCAAGGG-3′), Human DDX18(rDDX18-1: 5′-GATTGTTCAGTATGACCCTCCG-3′; rDDX18-2: 5′-CATGCCCTCTCCCATTTAGG-3′), Human FTSJ3(rFTSJ3-3: 5′-TCTCTGGATA GTGACCTGGATC-3′; rFTSJ3-4: 5′-ACTTCAGTAAGTCGCATACGC-3′), Human GAPDH(GAPDH-Fr: 5′-CTTTGTCAAGCTCATTTC CTGG-3′; GAPDH-Rr: 5′-TCTTCCTCTTGTGCTCTTGC-3′).

Immunoprecipitation Mass Spectrometry

C-terminal 3X-FLAG tagged expression clones of candidate ribosome biogenesis proteins were constructed via Gateway LR cloning (Invitrogen) of human ORF clones from the PlasmidID collection into a modified pcDNA3 vector (Invitrogen) followed by sequence verification. $3x10^6$ HEK293 cells were transfected with 5 μg of DNA of tagged genes and untransfected cells were used as control. FuGene6 (Roche) reagent in DMEM medium with 10% FBS and 1 U/ml of penicillin and streptomycin (Lonza) was used to transfect the cells for 24 hr. Cells were harvested after growing in the same medium with 10 U/ml of penicillin and streptomycin for an additional 24 hr. Cell lysis, FLAG immunoprecipitation (IP) on M2-agarose (Sigma; A2220), immuno-complex elution and digestions were performed according to the method of Dunham et al. (2011). Digested peptide mixtures (9 μl) were loaded onto a reverse phase micro-capillary pre-column (25-mm x 75-μm silica packed with 5-μm Luna C18 stationary phase; Phenomenex) and injected onto a micro-capillary analytical column (100-mm × 75-μm). Peptide separation was performed over 105 min with 5%–95% Acetonitrile (acidified with 0.1% formic acid) via an EASY-nLC system. Eluted peptides were directly sprayed into an Orbitrap Velos mass spectrometer (ThermoFisher Scientific) with collision activated dissociation using a nanospray ion source (Proxeon). 10 MS/MS data-dependent scans were acquired simultaneously with one high resolution (60,000) full scan mass spectrum. An exclusion list was enabled to exclude a maximum of 500 ions for 30 s. Acquired RAW files were extracted from the mass spectrometry data with

the extractms program and submitted for database searching using the SEQUEST search engine against a target-decoy UniProtKB/Swiss-Prot FASTA file. Search parameters were set to allow for one missed cleavage site, one variable modification of +16 for methionine oxidation and one fixed modification of +57 for cysteine carbamidomethylation using precursor ion tolerances of 3 m/z. After searching, peptide and protein hits were filtered using a 20 ppm tolerance for the precursor ion. We required 1% FDR for protein and peptide positive identifications.

Computational Analyses

MS Correlation Measures

Pearson Correlation Coefficient Score. Proteins belonging to the same multi-protein complex should co-elute across a biochemical fractionation, giving rise to similar elution profiles for those proteins. The similarity of elution profiles, represented as vectors containing the observed spectral counts for a protein in each fraction in a single experiment, was initially measured by Pearson correlation coefficient of the normalized elution profiles.

Each fractionation and mass spectrometry series identifies N proteins across M fractions. The raw data matrix is then an N by M matrix A where each $A(i,j)$ represents the number of MS/MS spectra observed to match protein i in fraction j. The normalized data matrix, B, converts numbers of peptides to frequencies, and is calculated as

$$B(i,j) = \frac{A(i,j)}{\sum_i A(i,j)}$$

A protein's normalized elution profile is represented by a row in this matrix, and the Pearson correlation coefficient was measured for each pair of proteins.

While the Pearson correlation coefficient is a good indicator of a co-complex relationship if both proteins are observed at high counts in the matrix, proteins observed at very low counts but found in the same fraction are often perfectly correlated but have poor predictive power (Figure S5).

To circumvent this artifact, we synthetically introduced noise into the raw data matrix and measured the extent to which noise affected the observed correlations and, by extension, the predictive power of correlation as it relates to protein complex membership. The observation of each protein in each fraction is modeled as a Poisson process, with lambda parameter assigned as the maximum likelihood estimate equal to the raw counts of protein i in fraction j (the $A(i,j)$ value). The noise term $1/M$ was added to the maximum likelihood estimate for each cell. The value $1/M$ was chosen on the basis that each protein was represented in the matrix by at least one peptide count, and the background probability for this should be evenly distributed across the M fractions. Thus the noise-added matrix $C = A + 1/M$, a constant. The MS experiment is re-run in silico by drawing randomly from Poisson($C(i,j)$) for each cell, then normalizing as above and calculating the Pearson correlations for each pair of proteins. This process was repeated 1,000 times, and the mean Pearson correlation for each pair was recorded. The noise term has the effect of giving every cell in the matrix a nonzero, albeit small, probability of "discovering" a protein count in that cell. The impact of this discovery on the correlation of that protein's elution profile with other normalized elution profiles is minimal for proteins observed at high counts and maximal for those observed with only one count across all fractions.

Weighted Cross-Correlation

In addition to the noise model correlation scores, a weighted cross correlation score was calculated for each pair of proteins in each experiment. We calculated the similarity of spectra profiles between each pair of proteins using a weighted cross correlation (WCC) approach (de Gelder et al., 2001), which was implemented in the R package wccsom (http://cran.r-project.org/web/packages/wccsom/index.html). The similarity value is between 0 and 1.

There are some advantages of this approach over other similarity measures, such as Pearson correlation coefficient. The WCC approach can take into account the relative shift between spectra profile patterns. In other words, given a protein, we can compare its spectra profile at a point/fraction with the profiles in that neighborhood of the corresponding point/fraction of another protein. Moreover, we can weight the different points in the neighborhood. In our calculation, we considered one point/fraction shift between spectra profile patterns and defined the weights based on a simple triangle function (http://mathworld.wolfram.com/TriangleFunction.html).

Machine Learning Methods

The noise-model correlations and weighted cross correlations of each pair of proteins observed in each of the seven cytoplasmic and eleven nuclear MS fractionation experiments were combined into matrices of protein pairs x 14 (cytoplasmic) or x 22 (nuclear) experimental observations. Missing data, where the pair of proteins were not both observed in a given experiment, were interpreted as zeros.

A gold standard reference set of positive and negative interactions was generated from the CORUM database of curated mammalian protein complexes. Human complexes consisting of 3 or more proteins were identified and filtered for those identified by mass spectrometry and related methods, removing those identified solely by, e.g., two-hybrid approaches, EMSA, and imaging techniques. Highly overlapping complexes (those with Simpson coefficient > 0.5) were merged, resulting in a reference set of 324 complexes comprised of 2,151 proteins. Each complex was then classified as "nuclear" and/or "cytoplasmic" based on the GO

Cellular Component annotation of its constituent proteins, resulting in 198 cytoplasmic and 190 nuclear complexes. These complexes were then randomly split into two groups, one for training pairwise co-complex protein-protein interactions in a machine learning framework and an independent set for optimizing final protein complex predictions from putative PPI. For PPI training, a reference positive interaction was defined as the case when two proteins were annotated to be in the same complex, and a reference negative interaction was defined where both proteins were in the annotated set but never appeared in the same complex. Although the CORUM complexes contain a large number of highly overlapping, redundant complex definitions, merging redundant complexes and reducing the complexes to unique pairwise interactions minimizes this source of bias. To further reduce bias, we omitted the largest complexes from the CORUM reference set (e.g., spliceosome, ribosome), which would otherwise account for a majority of reference PPI. Moreover, although our definition of negative interactions almost certainly contains some actual positives due to incomplete annotations, their effect is necessarily small, as negative interactions greatly outnumber positives. This renders our estimates of accuracy conservative, as some negatives will in fact be mislabeled. Our complete set of reference complexes is listed in Table S3.

The data were subjected to a variety of machine learning algorithms using the Weka suite of tools and assessed for accuracy and coverage. *Naive* Bayes and Logistic Regression classifiers were run using default parameters. Support Vector Machines (SVM) were applied using the SMO engine with a radial basis function kernel. The Random Forest implementation in Weka was too slow to use in an exploratory fashion but the Fast Random Forest re-implementation (http://code.google.com/p/fast-random-forest/) gave a significant performance boost and yielded the best results, as judged by cross-validated recall-precision analysis.

Incorporation of Genomic and Proteomic Evidence

Genomic and proteomic evidence were assembled from the HumanNet functional gene interaction network (Lee et al., 2011). HumanNet integrates a wide array of alternate data types across both human cell lines and model organism experiments into a log likelihood score indicating the strength of evidence suggesting that a given pair of genes operates in the same biological process. We considered only selected lines of evidence from HumanNet, excluding data derived from human experimental and computational prediction of protein-protein interactions, in order to minimize circularity that might bias predictions of PPIs. In all, protein-protein linkages from 17 lines of evidence were individually added to the classifier as independent features, with missing values set to zero. Table S6 lists the data types included in this study.

The nuclear data set thus comprised 41 quantitative features for each protein pair: 11 MS data sets measured by noise-model correlation, and again by weighted cross-correlation; the 17 features from HumanNet; a Co-Evolution score (Clark et al., 2011; Tillier and Charlebois, 2009) measuring correlated evolutionary rates; and a Co-Apex score measuring the number of MS experiments in which both proteins showed maximum (modal) abundance in the same fraction. Likewise, the cytoplasmic data set consisted of 33 features per pair: 14 MS and 19 other.

We used a greedy stepwise feature selection algorithm, implemented in Weka, to rank features and selected only the most informative ones, with the specific goal of choosing the single best correlation metric for each particular MS data set. It was observed that, after the first of the large-scale repeat MS experiments was folded into the classifier, the second repeat added little information and ranked poorly. To rescue these data, we merged the four largest repeats by addition and recalculated the noise model and weighted cross correlation scores for these four data sets. Performing feature selection on these data yielded 22 top-performing, non-duplicated features for the cytoplasmic data and 25 features for the nuclear data (Table S2). Predictions were generated for these sets using the Fast Random Forest classifier in Weka and a combined score was generated for each pair by taking one minus the product of one minus the posterior probability of the pair interacting, as predicted by the classifier. For pairs that appeared in only one data set, that data set's posterior probability was used. Applying the classifier to all pairs which had a correlation measure greater than 0.5 in any one MS data set yielded 817,179 protein pairs, of which 48,915 had posterior probability > = 0.5. Notably, incorporation of the complementary genomic evidence boosted the recall of PPI beyond that from the mass spectrometry evidence alone, across a wide range of predictive precision, e.g., increasing recall by ~20% at a cumulative precision of 0.7. The improvement shown by the final version of the data is shown in the main text in Figure 2C.

Denoising the Inferred Protein-Protein Interactions

We developed a procedure that exploits the network topology and protein co-localization information in order to further reduce the amount of noise in the inferred protein-protein interaction network and to filter it prior to discovering protein complexes.

We first delete the connections in the interaction network for which there is little evidence according to the network topology. The rationale here is that if two proteins belong to the same complex, they should be well connected to each other through many short paths in the graph. Diffusion methods over random graphs have previously been employed to quantify the amount of connectivity existing between two nodes in a graph (Coifman et al., 2005; Paccanaro et al., 2006).

Here we use a multiple-step diffusion which calculates the connectivity between proteins *i* and *j* as the *(i,j)* element of the matrix:

$$e^{\lambda \cdot M} - \lambda \cdot M$$

where *M* is the 5,549 × 5,549 matrix whose entries are the output of the random forest classifiers, and λ is the inverse of the maximal eigenvalue of M. Edges with diffusion values lower than *5E-05* are then deleted from the original graph. We shall indicate this new network with D.

Second, we calibrate the resulting graph using protein co-localization information.

To do this we combine the output of the previous step with the GO-CC (Harris et al., 2004) normalized semantic similarity scores with the assumption that they are independent. The rationale here is that two proteins located in different cellular locations should not interact. The final score for each link is thus given by:

$$1 - (1 - D(i,j)) \cdot \left(1 - \frac{Sim(i,j)}{MS}\right)$$

where *Sim(i,j)* is the maximum of the pairwise similarities between the two groups of GO-CC terms to which protein i and protein j are annotated, and MS is the maximum value among all the semantic similarity scores. In our calculations, for the semantic similarities we used an improved version of the Resnik semantic similarity measure (Resnik, 1999) that we have recently proposed (Yang et al., 2012) and is able to take into account the ontology beneath the GO terms and to model uncertainty.

Note that, among 5,549 proteins, there are 1,790 proteins that are not annotated in GO-CC. Therefore for these proteins we simply used D (output of the first step), as this (second) step cannot be applied to unannotated proteins. When considering the GO-CC annotation we discarded those with evidence codes NR, IEA, and ND.

Scores below a threshold of 0.55 were set to zero. The resulting denoised Protein-Protein Interactions graph contains 13,993 interactions (3,006 proteins) at an estimated 21.5% FDR. The effectiveness of the denoising procedure can be seen in a precision-recall curve for the network after denoising obtained by varying the threshold over the network weights and using as gold standard the CORUM database of curated mammalian protein complexes described earlier (Figure 2F).

Clustering of the Denoised PPI Network to Discover Protein Complexes

Protein complexes appear as densely connected regions within the de-noised interaction network. Because a protein may belong to multiple complexes, these densely connected regions may overlap. To elucidate such overlapping sets in our network, we used an algorithm that we have recently proposed, named ClusterONE (Clustering with Overlapping Neighborhood Expansion) (Nepusz et al., 2012). ClusterONE finds complexes by growing multiple clusters from seed proteins, independently of each other. The growth of a putative complex is governed by a greedy rule that tries to maximize the cohesiveness of the complex. The cohesiveness of a complex *C* is defined as follows:

$$\frac{w_{in}}{w_{in} + w_{out} + p|C|}$$

where W_{in} is the total weight of connections within *C*, W_{out} is the total weight of interactions connecting the complex with the rest of the network and $|C|$ is the size of the complex. *p* is a penalty constant that accounts for the possibility of uncharted connections in the network as it assumes *p* extra external connections for the complex for every protein involved. In each step of the growth process, we add a new adjacent protein to the complex or remove an already added protein in a way that yields the maximal increase in cohesiveness. The growth process stops when it is not possible to increase the cohesiveness further. At this stage, the cluster is declared a protein complex candidate if its density is above a given density threshold *d*, and the growth process restarts from a different seed. The first seed is the protein with the largest total weight on its incident connections (*i.e.*, the protein with the most confident set of interactions), and subsequent seeds are always selected in a similar manner but excluding proteins that have already been added to some protein complex candidate. Because the growth processes are independent of each other, the calculated complexes may overlap. More details on ClusterONE can be found in Nepusz et al. (2012). The algorithm has two main parameters: the penalty *p* and the density threshold *d*. The settings for these parameters were chosen to yield the highest Maximum Matching Ratio (Nepusz et al., 2012) on the cluster-training complex subset (see above). These were p = 2.9 and *d* = 0.4 and used to derive the final set of complexes.

To evaluate the overlap of the predicted complexes with the CORUM complexes, we calculated: (1) the number of CORUM complexes matching at least one predicted complex by a matching score greater than 0.25 (matching score = size of intersection squared, divided by the product of the two complexes sizes, as defined by (Bader and Hogue, 2003), (2) the Maximum Matching Ratio, (Nepusz et al., 2012), calculated by matching each predicted complex to at most one reference complex and vice versa, while maximizing the total matching score between them (with the theoretical maximum of 1.0 considered as a perfect match), (3) geometric accuracy as defined by (Brohée and van Helden, 2006) (square-root of the product of positive predictive value and clustering-wise sensitivity). The predicted complexes showed better correspondence with the CORUM catalog of reference human protein complexes than the results of other popular methods, including MCODE, MCL, CMC and RNSC (see Table S5). Applying ClusterONE to our denoised network, we obtained a set of 771 complexes. We then further filtered this set using the same procedure that we had applied to the CORUM set, which combined complexes sharing subunits (Simpson coefficient >0.5 between complexes). This produced our final set of 622 protein complexes.

Enrichment Analysis of Protein Pairs with Shared Annotations

To evaluate interacting and co-complexed protein pairs, we collected the following large-scale sets of protein-protein interactions: 1,991 co-complex interactions related to chromosome segregation (Hutchins et al., 2010); 17,775 "co-regulator" interactions identified through affinity purification and mass spectrometry-based methods (Malovannaya et al., 2011); and 209,913 interactions from

a *D. melanogaster* co-complex interaction network (Guruharsha et al., 2011). In addition, we collected the following sets of gene annotations: three available sets of 1,023, 3,563, and 114,477 human disease-gene associations (Becker et al., 2004; Hamosh et al., 2005; UniProt Consortium, 2011), 2,065 gene-mitotic phenotype associations (Hutchins et al., 2010; Neumann et al., 2010), curated sets of 74,250 mouse, 86,383 yeast, and 27,065 worm gene-phenotype associations assembled in (McGary et al., 2010), upstream transcription factor regulatory motifs for 265,270 genes (Xie et al., 2005), and a set of 869 essential genes collected from (Amsterdam et al., 2004; Blake et al., 2011; Harborth et al., 2001; Kittler et al., 2004; Silva et al., 2008).

We tested whether protein interaction partners are enriched for having common functional or phenotypic associations. That is, are protein pairs which are predicted to interact significantly more likely to share annotations? For each annotation set, we calculated the total number of protein pairs sharing annotations in the space of all possible pairs formed from the background set of annotated proteins detectable through our experimental procedures. We compared this ''expected'' fraction of pairs with shared annotations with the ''observed'' fraction of interaction partners with shared annotations. To measure the significance of the observed fraction, we obtained a *p*-value from the following hypergeometric test:

$$p(x \geq k) = \sum_{x=k}^{\min(n,m)} \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}, \quad (1)$$

where *N* is the number of possible annotated pairs, *m* is the number of possible pairs with shared annotation, *n* is the number of annotated interaction partners, and *k* is the number of interaction partners with shared annotation. In the case of testing for essentiality enrichment, we used the complete set of possible proteins pairs. Enrichments were additionally confirmed (data not shown) with two empirical *p*-values by calculating shared annotation fractions from 10,000 random trials, in which we (1) drew random protein partners from the background protein set and (2) shuffled the protein labels on the predicted protein interaction map. Lastly, we repeated the analysis for protein edges implied in our predicted protein cluster sets.

Tissue/Cell Line Specificity of Protein Complexes

It is important to note that HeLa cells were sampled in our profiling pipeline much more deeply than HEK, for which only nuclear fractionations were performed. Nevertheless, we examined the abundance of the interacting human proteins in HEK293 and HeLa cell lines on the basis of publicly available next-gen RNA sequencing data for both HeLa versus HEK293, well aware of the fact that mRNA expression levels may not necessarily reflect protein abundance. Considering all IEX MS experiments, HeLa proteins are discovered at slightly higher rates than those expressed at the same level in HEK293 (Figure S2B). We can clearly distinguish the few proteins that show differential tissue expression e.g., unique to one cell line. Among proteins assigned to complexes, only 82 show HeLa-specific expression and 11 HEK-specific expression (i.e., difference in Log2 (fpkm) expression > 2), yet these proteins show no preferential assortment into tissue-specific complexes.

The distribution of potentially tissue-specific proteins in complexes may reflect possible false positives arising from our analysis but is readily explained as a consequence of the false negative rate of protein detection, due to under-sampling by LC-MS. Hence, we directly examined the reproducibility of our fractionation/mass spectrometry data across biological replicates of the two cell lines, comparing MS1 intensities versus MS2 spectral counting as alternate methods of quantification. Moreover, it is worth noting that we find no evidence for stronger sampling biases in either proteome beyond what is to be expected for mass spectrometry in general. At the level of predicted PPI (which are derived from multiple biochemical fractions), we find that differences in the proteomic measurements generated for the two cell lines (again, in which HeLa was sampled far more extensively, particularly with regards to cytoplasmic extracts) lie within the variance actually observed between biological replicates of the same cell line (Figures S1 and S2).

The conclusion that the complexes we report are likely ubiquitous is supported by the expression of protein complex subunits across different tissues. For example, the Mann group surveyed the proteomes of 11 cancer cell lines; proteins in our complexes are generally found in all 11 lines (Figure S3A). Moreover, across 16 healthy human tissues for which RNA-seq data is available (EBI accession number E-MTAB-513), we find our complexed proteins to be highly and invariantly expressed (Figure S3B). Across 17,927 confirmed protein-coding genes detected in any of the 16 tissues, the median standard deviation of gene expression is 1.30, while for the 11,325 genes detected in all 16 tissues (63% of the total) it is 0.90. The standard deviation of genes we assign to protein complexes is 0.73; among these proteins, 91% are detected in all 16 healthy tissues. Thus the protein complexes described here exhibit largely invariant expression across the tissues sampled in the RNA-seq study.

Enrichment Analysis of Protein Clusters with Particular Phenotype Associations

We tested whether predicted protein clusters are enriched for particular human, mouse, or worm gene-phenotype associations. The significance of members of a cluster sharing a particular phenotype was determined by the hypergeometric probability, as above, where *N* is the number of annotated proteins in the background protein set, *m* is the number of proteins annotated with the queried phenotype, *n* is the number of annotated proteins in the cluster, and *k* is the number of proteins in the cluster annotated by the queried phenotype.

Cross-Validations with Curated Complexes in Public Databases and Independent Studies

We compared our network of complexes to curated complexes in 5 public databases, including CORUM (Ruepp et al., 2010), REACTOME (Haw et al., 2011), PINdb (Luc and Tempst, 2004), and HPRD (Prasad et al., 2009) databases, and specified complexes

within the GO cellular component category (Ashburner et al., 2000) to assess the agreement between our complexes and the literature. Statistically significant overlap between complexes was evaluated using the Fisher's exact test for hypergeometric distribution and the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) to correct for multiple testing (estimated false discovery rate ≤ 0.05), with a minimum of 2 shared subunits. Next, we validated putative new complexes (i.e., not curated in the above public repositories) through comparison with recently published independent co-affinity purification data (Guruharsha et al., 2011; Malovannaya et al., 2011). In particular, we accessed the recent human protein interaction results of Guruharsha et al. (2011). This group performed affinity-tag pull-down experiments for human proteins present in 41 of our complexes. Overall, of the 299 relevant human bait-prey interactions reported, 143 likewise occur within our complexes, representing a 47.8% validation rate. This agreement is comparable to the 63.8% validation rate they claim for their own complex predictions, and is probably an underestimate because they don't report all the proteins actually detected by mass spectrometry, but rather only human proteins with orthologs in their initial *Drosophila* PPI network. The matched clusters are reported in Table S3.

We also compared our complexes with the results of Malovannaya et al. (2011), which verified a total of 127 of our complexes (i.e., clusters show a Simpson matching coefficient > 0.5 between studies), including 42 (33%) of our complexes that are not curated in CORUM. These matched complexes are listed in Table S3. Taken together, these analyses represent a nearly 40% validation rate and strongly argue for the high fidelity of the mapped complexes.

Conservation of Complexes across Model Organisms

To examine to what extent human protein complexes identified in this study have known counterparts in yeast and fly, we considered the set of 720 multi-protein complexes in *S. cerevisiae* identified in a recent study (Babu et al., 2012) and the 556 complexes recently derived for *D. melanogaster* (Guruharsha et al., 2011). Both sets of complexes were identified using AP/MS techniques. Briefly, human complexes were converted into an ortholog representation by mapping, whenever possible, the components of each complex to their orthologs in yeast and fly, respectively. Using the ortholog representation of individual complexes, we then searched for the most statistically significant match between this representation and all known complexes from the corresponding organism. The process was also repeated in the opposite direction, mapping model-organism complexes onto the human collection in order to identify reciprocally best matches. Statistical significance was established using the Fisher's exact test for hypergeometric distribution and the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) to correct for multiple testing (estimated false discovery rate ≤ 0.05). Orthology relationships for human, yeast and fruit fly were derived from two well established sources: the InParanoid 7.0 (Ostlund et al., 2010) and Ensembl Compara (Vilella et al., 2009). The latter includes both the current Ensembl release 64 (ftp://ftp.ensembl.org/pub/release-64/mysql/ensembl_compara_64/) and Ensembl Genomes release 11 (ftp://ftp.ensemblgenomes.org/pub/pan_ensembl/release-11/mysql/ensembl_compara_pan_homology_11_64/). The Ensembl IDs from Compara were mapped using BioMart Perl API (http://www.biomart.org/martservice.html). In addition, we extended the human-to-yeast orthology map by matching human and yeast genes that share a common fly ortholog.

Coevolution

For the calculation of coevolution scores, we used the program MatrixMatchMaker (MMM) (Clark et al., 2011; Tillier and Charlebois, 2009). Orthologous protein sequence clusters were obtained from the OMA Database (Schneider et al., 2007) to obtain 204,689 eukaryotic groups that span 96 species, of which 20,800 contained human orthologs. The groups containing a human protein and at least 10 orthologous sequences were aligned using MAFFT (Katoh et al., 2005) and distance matrices were obtained by using protdist from PHYLIP (Felsenstein, 2005) with the PMB distance matrix (Veerassamy et al., 2003) to correct for multiple substitutions. We ran MMM in an all-by-all manner with a selected tolerance of 0.1 (10%) and chose to use taxon information such that only sequences from the same species could be matched.

Relative Evolutionary Rate

An average matrix was obtained by averaging the distance matrix entries over all of the OMA groups' matrices. We used the average matrix to compute the relative rate of an OMA group's evolution, as the ratio of its rate (average distance to the human ortholog) over the average matrix's rate for the same subset of species pairs. Values greater than 1 are proteins that are evolving faster than average, whereas values less than one indicate more slowly evolving proteins.

Evolutionary Age

The distribution of species present in the OMA orthologous groups determined the ancestral node in the phylogenetic tree of all eukaryotic species. The evolutionary distance from the human sequence to this last common ancestral node was then calculated and, in the case of complexes, averaged over the proteins in the complex. This gives an approximate evolutionary origin of the human orthologs.

Interaction Database and PPI Orthology

All OMA proteins were assigned ROGiDs based on their amino acid sequence. These IDs were then used to identify the known physical (or inferred by the author) protein-protein interactions from the iRefIndex database (Razick et al., 2008), which combines protein interaction data from multiple public databases: BIND, BioGRID, CORUM, DIP, HPRD, IntAct, MINT, MPact, MPPI and OPHID. Human protein interaction data were also downloaded from most of these public databases and some other online available resources independently. These databases / resources included BioGRID (Stark et al., 2011), DIP (Salwinski et al., 2004), MINT (Ceol et al., 2010), HPRD (Prasad et al., 2009), INTACT (Aranda et al., 2010), NCBI Gene (http://www.ncbi.nlm.nih.gov/gene/), CORUM (Ruepp et al., 2010) and the Human interactome database (Rual et al., 2005). Orthology of the PPIs was then determined using the species distribution of the OMA groups.

Approximation of Subunit Stoichiometries

The relative stoichiometries of interacting proteins were approximated from their associated mass spectral MS/MS counts as follows: For each pair of interacting proteins, we considered all biochemical fractions in which both proteins were observed, and calculated relative stoichiometries for interacting protein pairs observed together in at least 10 fractions. Their relative stoichiometry was estimated as the median (across the fractions) of the ratios of their MS/MS spectral counts divided by their expected ratios of spectral counts given the proteins' differences in numbers of potential tryptic peptides. This was calculated as:

$$Stoichiometry = median\left(\frac{c_{1,j}/e_1}{c_{2,j}/e_2}\right)$$

where $c_{1,j}$ and $c_{2,j}$ are the spectral counts of protein 1 and protein 2 in fraction j (out of n), and e_1 and e_2 are the numbers of potential tryptic peptides for proteins 1 and 2, respectively, calculated using the same parameters as in the initial identification of proteins from the raw mass spectrometry data (e.g., considering up to one missing tryptic cleavage and employing the same spectral lookup database). Stoichiometries estimated by this approach between ribosomal subunits and between core proteasomal subunits were consistent with the expected 1:1 ratios, as shown in Figure 7.

Evaluating Potential Bias

We evaluated our final complexes for possible biases toward hydrophobic or low abundant proteins, underrepresented organelles, and complex size–considerations that address some of the technical limitations of our approach. By design, insoluble membrane-associated (hydrophobic) protein complexes were largely missed in this study. Consistent with other proteomics studies, our data are biased toward highly expressed genes (Figure S2B). Our protein complexes are preferentially enriched for water-soluble nuclear and cytosolic proteins (Benjamini-corrected $p \leq 10^{-52}$ and $p \leq 10^{-12}$, respectively), which nevertheless cover a wide spectrum of biological functions (as judged by enrichment for diverse functional annotation terms).

We also compared both the isoelectric point (pI) and subunit memberships of our predicted protein complexes versus those reported in the CORUM database. To this end, we first minimized the inflated number of redundant protein complexes in CORUM by merging complexes with similar annotated subunit compositions but reported by different authors. We then integrated protein complexes with Simpson coefficients > 0.5 to deduce a consolidated non-redundant set of 734 curated protein complexes ranging from 2 to 142 (spliceosome) annotated protein subunits per complex. As shown in Figure S4B, we do not observe significant bias toward negatively ($pI \leq 7$) or positively ($pI \geq 7$) charged protein complexes in our data set as compared to CORUM.

Our clustering strategy, ClusterONE, underweights small clusters of size 2 or 3 in an effort to control the false positive rate, resulting in a peak of clusters at size = 4 subunits as evident in Figure 3A in the main text. Despite this apparent bias, ClusterONE outperformed the competing clustering algorithms we tested against the independent benchmark set of reference complexes, as detailed above and summarized in Table S5. In practice, we find that most competing algorithms yield an exceptionally large number of small clusters, for which it is difficult to establish meaningful measures of accuracy. Nevertheless, although our informatic approach yields complexes with a biased size distribution, overall our complexes show demonstrably good performance against the reference sets noted in the text.

SUPPLEMENTAL REFERENCES

Amsterdam, A., Nissen, R.M., Sun, Z., Swindell, E.C., Farrington, S., and Hopkins, N. (2004). Identification of 315 genes essential for early zebrafish development. Proc. Natl. Acad. Sci. USA *101*, 12792–12797.

Andersen, J.S., Lyon, C.E., Fox, A.H., Leung, A.K., Lam, Y.W., Steen, H., Mann, M., and Lamond, A.I. (2002). Directed proteomic analysis of the human nucleolus. Curr. Biol. *12*, 1–11.

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A.T., Kerrien, S., Khadake, J., et al. (2010). The IntAct molecular interaction database in 2010. Nucleic Acids Res. *38* (*Database issue*), D525–D531.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

Bader, G.D., and Hogue, C.W. (2003). An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics *4*, 2.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc., B *57*, 289–300.

Blake, J.A., Bult, C.J., Kadin, J.A., Richardson, J.E., and Eppig, J.T.; Mouse Genome Database Group (2011). The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. Nucleic Acids Res. *39* (*Database issue*), D842–D848.

Brohée, S., and van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics *7*, 488.

Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L., and Cesareni, G. (2010). MINT, the molecular interaction database: 2009 update. Nucleic Acids Res. *38* (*Database issue*), D532–D539.

Clark, G.W., Dar, V.U., Bezginov, A., Yang, J.M., Charlebois, R.L., and Tillier, E.R. (2011). Using coevolution to predict protein-protein interactions. Methods Mol. Biol. *781*, 237–256.

Coifman, R.R., Lafon, S., Lee, A.B., Maggioni, M., Nadler, B., Warner, F., and Zucker, S.W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. Proc. Natl. Acad. Sci. USA *102*, 7426–7431.

de Gelder, R., Wehrens, R., and Hageman, J.A. (2001). A generalized expression for the similarity of spectra: application to powder diffraction pattern classification. J. Comput. Chem. *22*, 273–289.

Dignam, J.D., Lebovitz, R.M., and Roeder, R.G. (1983). Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. Nucleic Acids Res. *11*, 1475–1489.

Dunham, W.H., Larsen, B., Tate, S., Badillo, B.G., Goudreault, M., Tehami, Y., Kislinger, T., and Gingras, A.C. (2011). A cost-benefit analysis of multidimensional fractionation of affinity purification-mass spectrometry samples. Proteomics *11*, 2603–2612.

Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package) version 3.6. Department of Genomic Sciences, University of Washington.

Harborth, J., Elbashir, S.M., Bechert, K., Tuschl, T., and Weber, K. (2001). Identification of essential genes in cultured mammalian cells using small interfering RNAs. J. Cell Sci. *114*, 4557–4565.

Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., et al.; Gene Ontology Consortium (2004). The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. *32* (Database issue), D258–D261.

Havugimana, P.C., Wong, P., and Emili, A. (2006). Enhanced proteomic analysis by HPLC prefractionation. In Handbook of Pharmaceutical Biotechnology, S.C. Gad, ed. (Hoboken, NJ: John Wiley & Sons), pp. 1491–1501.

Haw, R.A., Croft, D., Yung, C.K., Ndegwa, N., D'Eustachio, P., Hermjakob, H., and Stein, L.D. (2011). The Reactome BioMart. Database (Oxford) *2011*, bar031.

Katoh, K., Kuma, K., Miyata, T., and Toh, H. (2005). Improvement in the accuracy of multiple sequence alignment program MAFFT. Genome Inform *16*, 22–33.

Kittler, R., Putz, G., Pelletier, L., Poser, I., Heninger, A.K., Drechsel, D., Fischer, S., Konstantinova, I., Habermann, B., Grabner, H., et al. (2004). An endoribonuclease-prepared siRNA screen in human cells identifies genes essential for cell division. Nature *432*, 1036–1040.

Luc, P.V., and Tempst, P. (2004). PINdb: a database of nuclear protein complexes from human and yeast. Bioinformatics *20*, 1413–1415.

McGary, K.L., Park, T.J., Woods, J.O., Cha, H.J., Wallingford, J.B., and Marcotte, E.M. (2010). Systematic discovery of nonobvious human disease models through orthologous phenotypes. Proc. Natl. Acad. Sci. USA *107*, 6544–6549.

Morin, R., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T., McDonald, H., Varhol, R., Jones, S., and Marra, M. (2008). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques *45*, 81–94.

Ostlund, G., Schmitt, T., Forslund, K., Köstler, T., Messina, D.N., Roopra, S., Frings, O., and Sonnhammer, E.L. (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. *38* (*Database issue*), D196–D203.

Paccanaro, A., Casbon, J.A., and Saqi, M.A. (2006). Spectral clustering of protein sequences. Nucleic Acids Res. *34*, 1571–1580.

Prasad, T.S., Kandasamy, K., and Pandey, A. (2009). Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. Methods Mol. Biol. *577*, 67–79.

Razick, S., Magklaras, G., and Donaldson, I.M. (2008). iRefIndex: a consolidated protein interaction database with provenance. BMC Bioinformatics *9*, 405.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. J. Artif. Intell. Res. *11*, 95–130.

Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., et al. (2005). Towards a proteome-scale map of the human protein-protein interaction network. Nature *437*, 1173–1178.

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. *32* (*Database issue*), D449–D451.

Schneider, A., Dessimoz, C., and Gonnet, G.H. (2007). OMA Browser—exploring orthologous relations across 352 complete genomes. Bioinformatics *23*, 2180–2182.

Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. Nature *455*, 58–63.

Silva, J.M., Marran, K., Parker, J.S., Silva, J., Golding, M., Schlabach, M.R., Elledge, S.J., Hannon, G.J., and Chang, K. (2008). Profiling essential genes in human mammary cells by multiplex RNAi screening. Science *319*, 617–620.

Stark, C., Breitkreutz, B.J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Van Auken, K., Wang, X., Shi, X., et al. (2011). The BioGRID Interaction Database: 2011 update. Nucleic Acids Res. *39* (*Database issue*), D698–D704.

Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., et al. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science *321*, 956–960.

Tabb, D.L., McDonald, W.H., and Yates, J.R., III (2002). DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. J. Proteome Res. *1*, 21–26.

Veerassamy, S., Smith, A., and Tillier, E.R. (2003). A transition probability model for amino acid substitutions from blocks. J. Comput. Biol. *10*, 997–1010.

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. *19*, 327–335.

Wiśniewski, J.R., Zougman, A., Nagaraj, N., and Mann, M. (2009). Universal sample preparation method for proteome analysis. Nat. Methods *6*, 359–362.

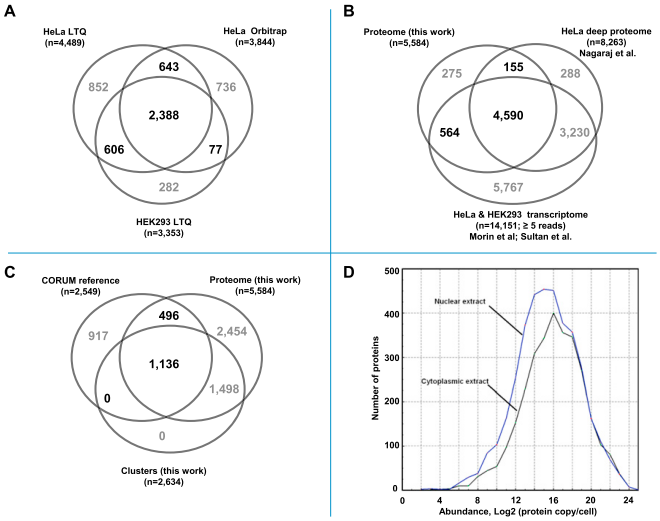


Figure S1. Assessment of LC-MS/MS Protein Detection Bias, Related to Figure 1 and Table S1

(A) Approximately 5% of proteins are unique to HEK cells (most likely to technical variations or sampling).

(B) Approximately 95% of the proteins identified in this study are supported by mRNA cognate transcript/or proteomic data produced with high resolution mass spectrometer (Nagaraj et al., 2011; Morin et al., 2008; Sultan et al., 2008).

(C) Proteins identified in this study covered 64% of the proteins present in the CORUM reference database.

(D) Deep fractionation allows to enrich and identify low abundance nuclear proteins by LC-MS/MS. Proteins abundances were estimated from recent study of HeLa Proteome by Mann group (Nagaraj et al., 2011).

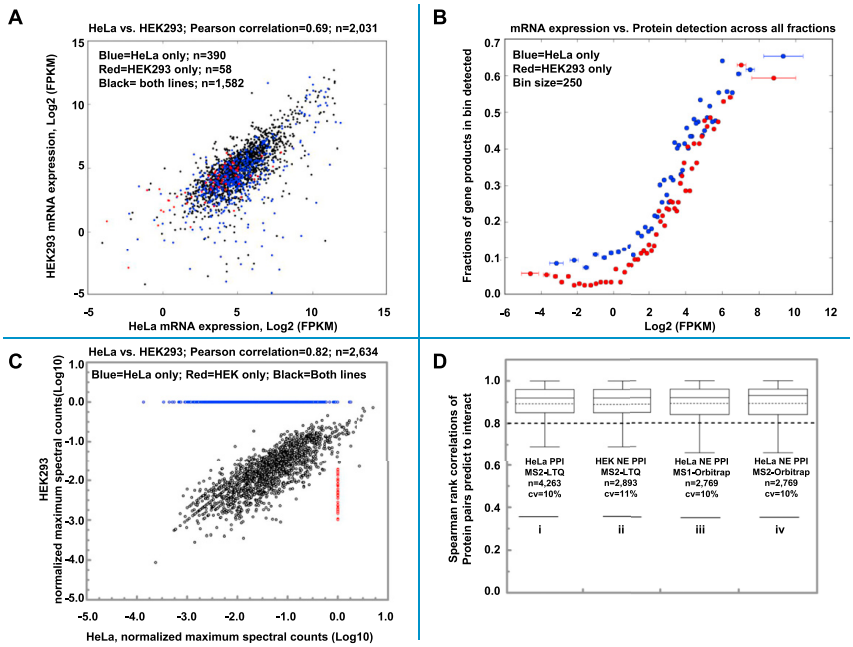


Figure S2. Comparison of HeLa and HEK Protein Profiles, Related to Figures 1 and 4 and Tables S1, S2, and S3

(A) Few proteins detected preferentially in HeLa or HEK293 cells have proportionally higher relative mRNA transcript levels in one of the two cell lines (Morin et al., 2008; Sultan et al., 2008); most show consistent transcript levels in both cell lines. Proteins detected in both cell lines are represented as black dots, and those detected only in HeLa or HEK cells are shown in blue or red, respectively.

(B) Gene products expressed in HeLa (blue) and HEK293 (red) cells (Morin et al., 2008; Sultan et al., 2008) were rank-ordered by mRNA-seq abundance level (log2(fpkm)) and binned (bin size = 250). For each bin, the fraction of gene products detected across all IEX fractionation experiments is plotted against the mean (+/− s.d.) expression of genes in the bin. Higher detection rate of HeLa proteins is consistent with deeper sampling of this cell line in our experiments.

(C) Positive correlation (r = 0.82) between HeLa (blue) and HEK (red) proteins assigned in our 622 complexes (2,634 proteins). For each protein in our set of 622 complexes, we retrieved its maximum spectral count across our 1,163 fraction and divided it by its length (i.e., number of amino acids). We then plotted the HEK versus HeLa after logarithmic transformation of the normalized spectral counts. Observed differences in protein detection, particular in HeLa, is mostly due to the protein detected in HeLa cytoplasmic extract.

(D) Box-and-whiskers quartile plots showing the high consistency (profile correlation > 0.8) of the co-fractionation data using different measures of protein abundance (MS2 spectral counts versus MS1 peptide intensities). Data reproducibility was calculated using the Spearman rank correlation coefficients of replicate profiles. Horizontal solid lines mark the minimum, first quartile, median, third quartile and maximum spearman correlation values; black dashed lines mark mean Spearman correlations. High-scoring interacting protein pairs show reproducible HeLa and HEK293 co-elution profiles measured on a linear ion-trap (i and ii, MS2 spectral counts for HeLa and HEK293, respectively) or a high precision Orbitrap instrument (iv, MS2 spectral counts; iii, MS1 peptide intensities based on MaxQuant).

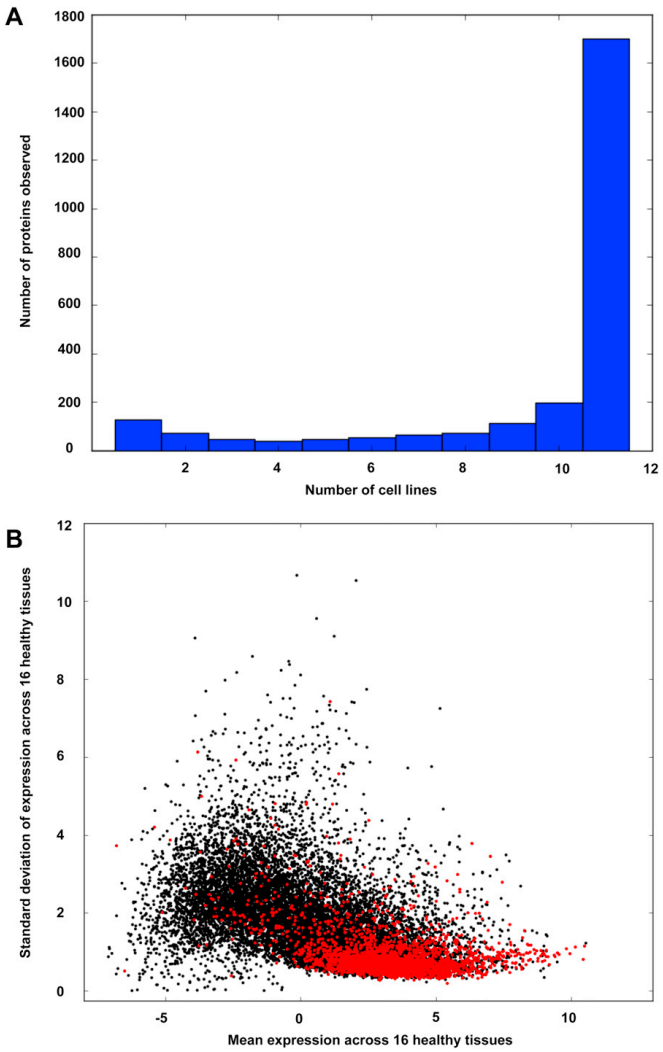


Figure S3. Tissue Expression of Proteins in Complexes, Related to Figure 6

(A) Histogram of number of cancer cell lines in which proteins assigned to our complexes were observed. Data from Mann group proteomic survey of 11 cancer cell lines (Geiger et al., 2012).

(B) Expression levels of RefSeq protein-coding genes across 16 healthy human tissues measured using the Illumina BodyMap 2.0 RNA-seq data (EBI accession E-MTAB-513). Here, mean expression (log2(fpkm)) across all tissues in which a gene product is observed is plotted against the standard deviation of expression: black, all genes; red, subunits assigned to protein complexes in this study. High mean and low variability of expression among protein complex components implies ubiquitous expression.

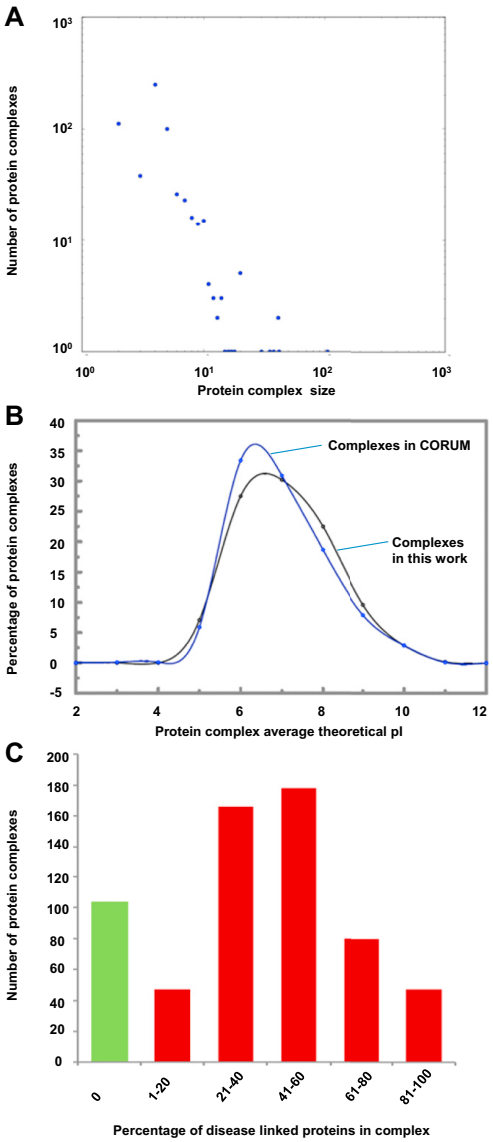


Figure S4. Physical and Biological Properties of our Predicted Human Protein Complexes, Related to Figure 4 and Table S3

(A) Size distribution of mapped protein complexes. The frequency distribution of the number of proteins per complex approximates an inverse power law.

(B) Evaluating bias in complexes. Theoretical pI for each individual protein was calculated using the open source "Compute pI/Mw" tool from the ExPASy (http://web.expasy.org/compute_pi/). To estimate the pI of the protein complex, theoretical pI for individual proteins in complex were averaged and rounded to integer values. Blue; complexes in CORUM reference. Black; complexes derived in this study.

(C) Distribution of annotated disease-associated proteins that are present in our compendium of 622 protein complexes.

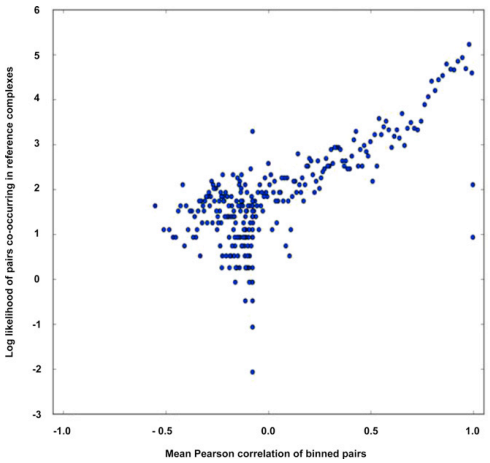


Figure S5. Pearson Correlation between Elution Profiles Breaks Down at High Correlations, Related to Figure 2B

Data from the cytoplasmic fraction of the sucrose gradient MS experiment were analyzed by ranking pairs according to the Pearson correlation coefficient of the normalized elution profiles (*x* axis), binning, and calculating for each bin the log likelihood of containing reference set co-complex protein pairs (*y* axis). Correlation coefficient is predictive of LLS score but breaks down as correlation approaches unity. This drop-off is caused by low-count proteins showing perfect correlations, and was compensated through the use of a Poisson weighted correlation test.

Appendix B

Tempo and Mode of Genome Evolution in a 50,000 Generation Experiment[1,2]

[1]Tenaillon, O [*et al.*, including Wu, GC]. Tempo and Mode of Genome Evolution in a 50,000 Generation Experiment. Nature, 536(7615):165170, August 2016.

[2]In the following work, I performed fitness experiments, and analyzed genomes, including synonymous and non-synonymous mutations, and intergenic mutations. In addition, I wrote data analysis tools for parsing the breseq output that aided in the analysis. I also edited and reviewed the manuscript prior to publication and approved the final version.

ARTICLE

doi:10.1038/nature18959

Tempo and mode of genome evolution in a 50,000-generation experiment

Olivier Tenaillon[1]*, Jeffrey E. Barrick[2,3]*, Noah Ribeck[3,4], Daniel E. Deatherage[2], Jeffrey L. Blanchard[5], Aurko Dasgupta[2]†, Gabriel C. Wu[2], Sébastien Wielgoss[6,7], Stéphane Cruveiller[8], Claudine Médigue[8], Dominique Schneider[7,9] & Richard E. Lenski[3,4]*

Adaptation by natural selection depends on the rates, effects and interactions of many mutations, making it difficult to determine what proportion of mutations in an evolving lineage are beneficial. Here we analysed 264 complete genomes from 12 *Escherichia coli* populations to characterize their dynamics over 50,000 generations. The populations that retained the ancestral mutation rate support a model in which most fixed mutations are beneficial, the fraction of beneficial mutations declines as fitness rises, and neutral mutations accumulate at a constant rate. We also compared these populations to mutation-accumulation lines evolved under a bottlenecking regime that minimizes selection. Nonsynonymous mutations, intergenic mutations, insertions and deletions are overrepresented in the long-term populations, further supporting the inference that most mutations that reached high frequency were favoured by selection. These results illuminate the shifting balance of forces that govern genome evolution in populations adapting to a new environment.

Comparative genomic studies have identified the molecular basis of adaptations including lactase permanence in humans[1], domestication of plants[2] and animals[3], and pathogenicity in bacteria[4]. Nevertheless, it is difficult to determine more generally what fraction of new mutations in an evolving lineage are beneficial. Answering this question is important for modelling sequence changes used in phylogenetic methods[5] and would inform debate about adaptive and non-adaptive modes of genome evolution[6,7].

The combination of experimental evolution and genome sequencing provides a way forward that has been used with viruses, bacteria, yeast and flies[8–13]. In a study of bacteria, the diversity of mutations involved in adaptation to high-temperature stress was studied by sequencing >100 lineages after a 2,000-generation experiment[10]. In another study, sequencing a series of clones from one population over 40,000 generations showed the trajectory of genome evolution[9]. However, a short-term experiment reveals only the early steps of adaptation, and it is difficult to distinguish adaptive 'driver' and non-adaptive 'passenger' mutations when only one population is examined. Beneficial mutations can also be identified by lineage tracking[14] and genetic reconstruction[15] experiments, but these approaches become impractical after an initial selective sweep or when mutations become too numerous over time, respectively.

To overcome these limitations, we analysed complete genomes of 264 clones from 12 populations across 50,000 generations of the long-term evolution experiment (LTEE) with *E. coli*[16,17]. These populations have evolved in a defined medium with scarce resources since 1988. Mean fitness measured in competition with their ancestor increased by ~70% in that time[17]. The LTEE is a model system for studying many fundamental evolutionary questions[9,15–23].

Genome-wide mutations and hypermutability

We sequenced the genomes of two clones from each population after 500, 1,000, 1,500, 2,000, 5,000, 10,000, 15,000, 20,000, 30,000, 40,000 and 50,000 generations using the Illumina platform (Supplementary Data 1). We called mutations, including structural variants, using the *breseq* pipeline[24,25]. In total, we found 14,572 point mutations; 500 insertions of insertion sequence (IS) elements; 726 deletions and 1,132 insertions each ≤ 50 base pairs (bp) (small indels); and 267 deletions and 45 duplications each >50 bp (large indels). After 50,000 generations, average genome length declined by 63 kb (~1.4%) relative to the ancestor (Extended Data Fig. 1). Mutations were not distributed uniformly across the populations. Instead, six populations (Ara−1, Ara−2, Ara−3, Ara−4, Ara+3 and Ara+6) had 96.5% of the point mutations, having evolved hypermutable phenotypes caused by mutations that affect DNA repair or removal of oxidized nucleotides[18,20]. Figure 1a shows the trajectories for the total mutations in all 12 populations; Fig. 1b is rescaled for better resolution of those that did not become point-mutation mutators. Hypermutability tended to decline over time as the load of deleterious mutations favoured antimutator alleles[20]. All four populations that were hypermutable at 10,000 generations accumulated synonymous substitutions (a proxy for the underlying point-mutation rate) between generations 40,000 and 50,000 at much lower rates than from 10,000 to 20,000 generations (Extended Data Fig. 2).

Increased numbers of IS elements can also cause hypermutability[26], with higher rates not only of transpositions but also deletions and duplications through homologous recombination. In population Ara+1, 31.8% of all mutations up to 50,000 generations were IS*150* insertions, compared with 12.3% for the other populations that never evolved elevated point-mutation rates. This mode of hypermutability arose early in Ara+1; IS*150* insertions are overrepresented in each Ara+1 clone from 5,000 generations onwards when compared individually to all other non-mutator clones from the same generation (Fisher's exact test with Bonferroni correction, $P < 0.05$). Two clones from other populations were also IS*150* hypermutators by this test: 38.7% of the mutations in

[1]IAME, UMR 1137, INSERM, Université Paris Diderot, Sorbonne Paris Cité, F-75018 Paris, France. [2]Department of Molecular Biosciences, Institute for Cellular and Molecular Biology, Center for Systems and Synthetic Biology, Center for Computational Biology and Bioinformatics, The University of Texas at Austin, Austin, Texas 78712, USA. [3]BEACON Center for the Study of Evolution in Action, Michigan State University, East Lansing, Michigan 48824, USA. [4]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, Michigan 48824, USA. [5]Department of Biology, University of Massachusetts, Amherst, Massachusetts 01003, USA. [6]Institute of Integrative Biology, ETH Zürich, Universitätstrasse 16, Zürich 8092, Switzerland. [7]Université Grenoble Alpes, Laboratoire Technologies de l'Ingénierie Médicale et de la Complexité — Informatique, Mathématiques et Applications (TIMC-IMAG), F-38000 Grenoble, France. [8]UMR 8030, CNRS, Université Évry-Val-d'Essonne, CEA, Institut de Génomique, Laboratoire d'Analyses Bioinformatiques pour la Génomique et le Métabolisme, F-91000 Évry, France. [9]Centre National de la Recherche Scientifique, TIMC-IMAG, F-38000 Grenoble, France. †Present address: Department of Internal Medicine, Washington University School of Medicine, St Louis, Missouri 63110, USA.
*These authors contributed equally to this work.

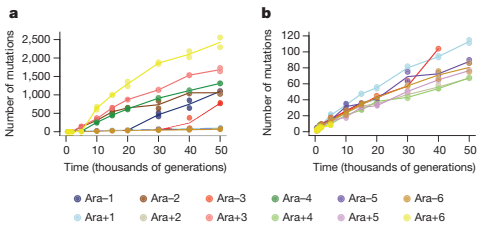


Figure 1 | Total number of mutations over time in the 12 LTEE populations. a, Total mutations in each population. **b**, Total mutations rescaled to reveal the trajectories for the six populations that did not become hypermutable for point mutations, and for the other six before they evolved hypermutability. Each symbol shows a sequenced genome; some points are hidden behind others. Each line passes through the average of the genomes from the same population and generation.

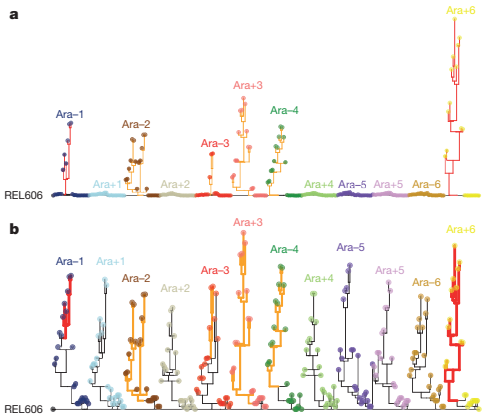


Figure 2 | Phylogenetic trees for LTEE populations. a, Phylogenies for 22 genomes from each population, based on point mutations. **b**, The same trees, except branches are rescaled as follows: branches for lineages with mismatch-repair defects are orange and shortened by a factor of 25; branches for *mutT* mutators are red and shortened by a factor of 50. Strain REL606 (on the left) is the ancestor. No early mutations are shared between any populations, confirming their independent evolution. Most populations have multiple basal lineages that reflect early diversification and extinction; some have deeply divergent lineages with sustained persistence, most notably Ara−2.

a 30,000-generation clone from Ara−5 and 31.7% of the mutations in a 40,000-generation clone from Ara−3 were IS*150* insertions. The aberrant Ara−5 clone shares only one mutation with other sequenced Ara−5 clones, indicating early divergence; it does not share point mutations with any other population, excluding cross-contamination. The emergence of these various mutator types shows that evolution can alter the production of genetic diversity[20,27], which in turn changes the tempo and mode of genome evolution.

Population phylogenies

Figure 2a shows phylogenetic trees constructed using point mutations for each population; Fig. 2b shows the trees with branches rescaled after mutators evolved. Some populations—including Ara−2, which became hypermutable early, and Ara−6, which never did—harbour lineages that coexisted for tens of thousands of generations. Some others—including Ara−4, which became hypermutable, and Ara+2, which did not—are more linear in structure, without deep branches among the sequenced clones. Deep branches were probably supported by the diversity-promoting effects of negative-frequency-dependent interactions, as shown in the Ara−2 population[22,23]. Sequencing whole-population samples would provide more detailed information on within-population diversity[11,12].

Dynamics of genome evolution

The accumulation of point mutations increased greatly in hypermutable populations[9,19,20], potentially overwhelming the genomic signature of adaptation. Although mutator lineages may experience higher rates of fitness improvement[17,27], the effect is usually small owing to clonal interference between competing beneficial mutations[28,29] and the increased load of deleterious mutations[20,30]. Therefore, beneficial mutations become harder to detect in a sea of unselected mutations in mutator lineages. To understand better the dynamic coupling between adaptation and genome evolution, we first analysed the populations that retained the ancestral mutation rate up to 50,000 generations and the others before they became point-mutation or IS*150* mutators.

It was previously found[17] that the mean-fitness trajectory of the LTEE is well described by a power-law relation, in which log fitness increases linearly with log time. Moreover, the power law accurately predicts fitness to 50,000 generations using data from only the first 5,000 generations. It was shown that a population-dynamical model that incorporates two phenomena known to be important in the LTEE—clonal interference[29,31] and diminishing-returns epistasis[15,29]—generates a power-law relation. This model in turn predicts that the number of beneficial mutations should increase with the square root of time[17]. However, not all mutations that accumulate are beneficial; neutral and nearly neutral mutations can spread by recurring mutation, random drift, and hitchhiking[32–34]. Selective sweeps will purge some neutral mutations but cause others to increase; overall, the expected number of neutral mutations should increase linearly with time[35].

To test these predictions, we fit three models to the trajectory for the total number of mutations in the non-mutator and premutator lineages:

$$m = at$$

$$m = b\sqrt{t}$$

$$m = at + b\sqrt{t}$$

where m is the number of mutations, t is time (generations), and a and b govern the genome-wide rates of accumulation of neutral and beneficial mutations, respectively (Fig. 3). (Extended Data Fig. 3 shows the models fit to each population separately.) Using the Akaike information criterion (AIC), the two-parameter model fits the data much better than those with only the linear (ΔAIC = −77.7) or square-root (ΔAIC = −99.7) terms. Because the one-parameter models are nested within the two-parameter model, we can also assess the significance of adding the second parameter; P values are 7.5×10^{-5} and 5.2×10^{-7} relative to the linear and square-root models, respectively. The trajectory for genome evolution thus shows signatures of both adaptive and non-adaptive changes. However, the model that predicts the square-root trajectory of beneficial substitutions makes various assumptions (for example, about the form of epistasis), and both the predicted and observed trajectories have statistical uncertainties. (Extended Data Fig. 4 shows the uncertainty in estimating a and b from the observed trajectory.) Therefore, we examined additional evidence to shed light on the proportion and identity of beneficial mutations.

Evidence for beneficial mutations

We sought to understand what proportion of the genomic changes in the non-mutator populations was adaptive, and how that proportion changed over time. One line of evidence derives from the expectation that synonymous substitutions—point mutations in protein-coding genes that do not affect the amino-acid sequence—are neutral and should therefore accumulate at a rate equal to the underlying

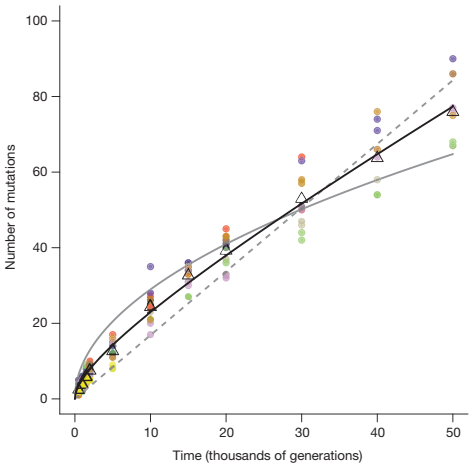


Figure 3 | Alternative models fit to the trajectory of genome evolution. Each symbol shows total mutations in a clone from five populations that never became mutators and seven before point mutation or IS*150* hypermutability evolved. Colours are the same as in Fig. 1; open triangles indicate grand means. Dashed grey line shows the best fit to the linear model, $m = at$. Solid grey curve shows the fit to the square-root model, $m = b\sqrt{t}$. Black curve is fit to the composite model, $m = at + b\sqrt{t}$, where $a = 0.000944$ and $b = 0.134856$. See text for statistical analysis.

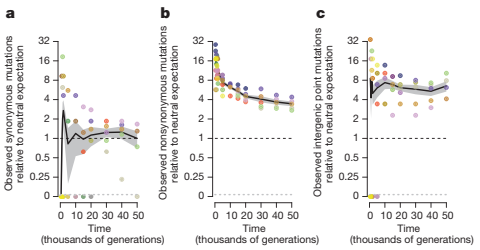


Figure 4 | Trajectories for synonymous, nonsynonymous and intergenic point mutations. a, Synonymous mutations, scaled so that the mean of five non-mutator populations (excluding point mutation and IS*150* hypermutators) is unity at 50,000 generations. **b**, Nonsynonymous mutations, scaled using the same rate as synonymous mutations after adjusting for sites at risk for both classes. **c**, Intergenic point mutations, scaled using the same rate as synonymous mutations after adjusting for sites at risk. Each symbol shows the mean for sequenced genomes from a non-mutator or premutator lineage. Colours are as in Fig. 1. Note the discontinuous scale; populations with zero mutations are plotted below. Black lines connect grand means; shading shows standard errors calculated from replicate populations.

mutation rate[20,35]. This expectation is not strictly true owing to selection on codon usage, RNA folding, and other effects, but it is generally thought that such selection is extremely weak, affects only a small fraction of sites at risk for synonymous mutations, or both[36,37]. We calculate whether nonsynonymous and intergenic point mutations are found in excess relative to synonymous mutations, given the number of sites at risk for each class. Figure 4a shows the number of synonymous mutations in non-mutator and premutator populations, scaled so the mean at 50,000 generations is unity. As expected, synonymous mutations accumulated at an approximately constant rate (Extended Data Fig. 5). Figure 4b shows the number of nonsynonymous mutations relative to the neutral expectation based on synonymous mutations. Nonsynonymous mutations accumulated ~17.1 times faster than synonymous ones during the first 500 generations and ~3.4 times faster over 50,000 generations. Nonsynonymous mutations continued to accumulate at over twice the rate of synonymous mutations in the later generations (Extended Data Fig. 6), implying that most nonsynonymous mutations that reached high frequency were beneficial even after so long in a constant environment. The same approach applied to intergenic point mutations (Fig. 4c) also reveals a large excess relative to synonymous mutations, although the number of events is smaller and the uncertainty greater. This result implicates adaptive changes in noncoding regions that presumably affect the binding sites for regulatory proteins[38–40].

Synonymous mutations provide an internal benchmark for nonsynonymous and intergenic point mutations. However, synonymous mutations are not directly informative for understanding how selection affects the accumulation of indels that comprise almost half the mutations in non-mutator clones at 50,000 generations (Extended Data Fig. 7). To estimate the proportion of beneficial changes for other types of mutation, we compare the LTEE and a mutation accumulation experiment (MAE) in which 15 lines were propagated via repeated single-cell bottlenecks[41]. Such bottlenecks eliminate the variation needed for natural selection, so that all types of mutations accumulate at the rates at which they happen, regardless of fitness effects, except for lethal or highly deleterious mutations that preclude cells from making colonies used to propagate lines[29]. MAE lines thus provide an external baseline for distinguishing beneficial and non-beneficial mutations. In fact, because more unselected mutations are deleterious than beneficial, MAE lines are expected to lose fitness over time, which they did (Extended Data Fig. 8).

To quantify the relative rates for all types of mutations in the absence of selection, we sequenced clones from the MAE lines after 550 daily bottlenecks (Supplementary Data 1). Consistent with the random accumulation of mutations, the number of nonsynonymous (including nonsense) mutations was similar to the expectation based on synonymous mutations (117 observed, 105.02 expected); the resulting ratio of 1.11 is well within the 95% confidence interval (0.70–1.50) obtained by a randomization test. Also, there was no among-line variation in total mutations ($\chi^2 = 5.46$, degrees of freedom (df) = 14, $P = 0.978$). We can therefore reasonably use the MAE lines to estimate relative rates of different types of mutations, with synonymous ones providing a benchmark largely free of selection in both experiments. For example, LTEE population Ara−1 had 21 nonsynonymous mutations at 20,000 generations and the expected number of synonymous mutations based on the average non-mutator population was 1.08 (Extended Data Fig. 5); the 15 MAE lines in total had 117 nonsynonymous and 39 synonymous mutations; thus, the ratio of observed mutations to the neutral expectation is (21/1.08)/(117/39) = 6.5. These ratios show that all major classes of mutations—including various indels—are substantially overrepresented in the LTEE relative to the MAE (Extended Data Fig. 9), implying that many mutations in each class were adaptive during the LTEE.

Parallel evolution at many gene loci

Parallel evolution occurs when similar changes arise independently in multiple lineages, and it is often used to discover putative targets of selection[4,8,10–13,21]. Genetic parallelism can be studied at the level of DNA sequence, affected genes, or integrated functions. Parallelism at the nucleotide level tends to be rare because different mutations in a gene often produce similar benefits[4,10–12,21], although there are exceptions[8]. Parallelism at a functional level requires detailed understanding that may be unavailable, and it is difficult to interpret when there are many mutations. We therefore examined parallelism at the gene level.

Table 1 | Protein-coding genes with the highest *G* scores

Gene	Length	Observed	Expected	*G*	Annotation
pykF	1,413	19	0.16	181	Pyruvate kinase
iclR	825	13	0.10	128	Transcriptional repressor, glyoxylate bypass
spoT	2,109	14	0.25	113	Stringent response
nadR	1,233	12	0.14	106	Bifunctional transcriptional repressor and NMN adenylyltransferase
hslU	1,332	11	0.15	94	Molecular chaperone and ATPase component of protease
yijC (also known as *fabR*)	705	7	0.08	62	Transcriptional repressor, fatty acid and phosphatidic acid pathway
topA	2,598	8	0.30	52	DNA topoisomerase I subunit
malT	2,706	8	0.31	52	Transcriptional activator, maltotriose-ATP-binding
mrdA	1,902	7	0.22	48	Transpeptidase in peptidoglycan synthesis
mreB	1,044	6	0.12	47	Longitudinal peptidoglycan synthesis
infB	2,673	7	0.31	44	Translation initiation factor IF-2
arcA	717	5	0.08	41	Response regulator in two-component system, anoxic redox control
argR	471	4	0.05	34	Repressor of arginine regulon
rplF	534	4	0.06	33	50S ribosomal subunit protein
mreC	1,104	4	0.13	28	Longitudinal peptidoglycan synthesis

Genes are ranked by *G* scores computed using observed independent nonsynonymous mutations relative to expected number given gene length (bp). Data are from populations with the ancestral point-mutation rate throughout and other populations before they evolved hypermutability.

We focused on lineages that retained the ancestral point-mutation rate (including clones from populations that later became hypermutable) because, as shown earlier, most mutations are drivers in those cases; we expect hypermutability to make the analysis less informative because many more mutations are passengers. We first calculated the expected number of nonsynonymous mutations for each single-copy protein-coding gene based on its length as a fraction of all such genes and the total number of nonsynonymous mutations in the relevant lineages (Supplementary Data 2). We computed *G* scores for goodness of fit between observed and expected values; the total score is 2,593.7. We compared that total with simulated data sets in which positions of mutations in the coding genome were randomized, and the observed total significantly exceeded the simulations (mean simulated $G = 1{,}933.7$, $Z = 25.5$, $P < 10^{-143}$). Fifty-seven genes had two or more mutations; these genes had 50.1% of the nonsynonymous mutations but constituted only 2.1% of the coding genome. (Only one gene had multiple synonymous changes.) Table 1 shows the 15 genes that contribute the most to the total *G* score. Several encode proteins with core metabolic or regulatory functions, including three involved in peptidoglycan synthesis.

We ran the same analysis for lineages that evolved hypermutability (Supplementary Data 3), and the randomization test indicates significant parallelism (G statistic $= 5{,}098.4$, mean simulated $G = 4{,}581.1$, $Z = 5.745$, $P < 10^{-8}$). As expected, however, the signal-to-noise ratio reflected in the significance level is much weaker than for the non-mutator lineages. Most genes with the highest scores in mutator lineages differ from those in non-mutators, in part because those genes often had beneficial mutations before hypermutability evolved.

Table 2 lists the 16 genes with the most deletions, duplications, insertions and intergenic point mutations in non-mutator lineages

Table 2 | Genes with the most mutations of other types

Genes	Mutations	Number	IS	MAE	Annotation
rbsD	Mostly large deletions	41	Yes	No	D-Ribose utilization; most deletions affect entire *rbs* operon
nupC	Various intergenic	19	Yes	Yes	Nucleoside transporter
iap	Mostly large indels	19	Yes	No	Alkaline-phosphatase isozyme conversion; most indels affect tens of adjacent genes including *rpoS*, which encodes stationary-phase σ factor
mokB	Various indels	17	Yes	Yes	Enables *hokB* toxin expression
yhgI/gntT	Intergenic point mutations	16	No	No	Gluconate transport
mokC	Various indels	15	Yes	Yes	Enables *hokC* toxin expression
ybcU (also known as *borD*)	Large indels	14	Yes	No	Indels affect this and adjacent remnants of DLP12 prophage
ECB_02013	Various indels	14	No	Yes	Indels affect this and adjacent remnants of P2-like prophage
ECB_02816 (also known as *kpsD*)	Various indels	14	Yes	No	Polysialic-acid transport protein precursor
acs/nrfA	Various intergenic	14	No	No	Acetyl-CoA synthase; nitrite reductase
hokE	Large indels	12	Yes	No	Toxin in plasmid-derived toxin–antitoxin system; most indels affect several adjacent genes involved in iron acquisition
ybeB/phpB	Various intergenic	11	Yes	No	Unknown functions, but adjacent to genes involved in cell-wall synthesis
ydiJ/ydiK	Various intergenic	11	No	No	Predicted FAD-linked oxidoreductase; putative inner membrane protein
ldrC	Various indels	10	Yes	Yes	Small toxic polypeptide
menC	IS insertions	10	Yes	Yes	Menaquinone biosynthesis
fimA	Mostly IS insertions	10	Yes	No	Component of fimbrial complex

Genes are ranked by total mutations excluding nonsynonymous and synonymous point mutations. When two genes are separated by a solidus, the affected sequence includes the intergenic region between them. IS column indicates whether the majority of mutations involve IS elements. MAE column indicates whether the same or nearly identical mutations occurred in one or more MAE lines. Data are from populations with the ancestral point-mutation rate throughout and others before they evolved hypermutability.

(Supplementary Data 2). For mutations that impact multiple genes, we show the most frequently affected gene (or adjacent pair when most events are intergenic). In 12 cases, the majority of the mutations were mediated by IS elements; these include insertions as well as deletions and duplications that appear to involve homologous recombination. In six cases (five with IS insertions), the same or nearly identical mutations occurred in one or more MAE lines, suggesting mutational hotspots. These changes may indicate high-frequency events, but recall that IS insertions and large indels are enriched in the LTEE relative to the MAE (Extended Data Fig. 9), implying that many are also beneficial. Indeed, the IS-mediated *rbsD* deletions occur at a high rate and are beneficial in the LTEE environment[42], and some IS-mediated mutations appear to be beneficial in other studies as well[43,44].

The parallelisms involving nonsynonymous substitutions and other mutations in the LTEE, coupled with their high rates of accumulation relative to the MAE, indicate that many observed mutations were drivers of adaptation. For indels, however, the specific target genes are difficult to identify owing to the multiplicity of genes affected and the potentially confounding effect of mutational hotspots.

Discussion

Adaptation by natural selection sits at the heart of phenotypic evolution. However, the random processes of spontaneous mutation and genetic drift often overwhelm and obscure genomic signatures of adaptation. We overcame this difficulty by analysing genomes from 12 bacterial populations that evolved for 50,000 generations under identical culture conditions. Even so, six populations evolved hypermutable phenotypes that increased point-mutation rates ~100-fold, and another evolved hypermutability caused by a transposable element. By focusing on populations that retained the ancestral mutation rate, we identified several key features of the tempo and mode of their genome evolution. First, a population-genetic model with two terms—one for beneficial drivers, the other for neutral hitchhikers—fits the dynamics much better than models without both terms. Second, the great majority of mutations observed during the early generations were beneficial drivers. Third, the proportion of observed mutations that were beneficial declined over time but remained substantial even after 50,000 generations. The second and third findings follow from the population-genetic model. Both are also strongly supported by the excess of nonsynonymous to synonymous substitutions in the LTEE and by the excess of several classes of mutations, including indels, in comparison to mutation-accumulation lines. Fourth, there was strong gene-level parallel evolution across the replicate LTEE populations.

Our analyses also show a contrast between the contributions of beneficial mutations to molecular evolution and to the fitness trajectory in a stable environment. In particular, beneficial mutations continued to constitute a large fraction of genetic changes throughout the 50,000 generations of the LTEE, whereas the resulting fitness gains were only a few per cent in the last 10,000 generations[17]. Beneficial mutations with very small selection coefficients are nonetheless visible to natural selection[17]. Hence, adaptation can remain a major driver of molecular evolution long after an environmental shift. Our experimental results thus support a selectionist view of molecular evolution, complementing indirect evidence based on comparative genomics in bacteria, *Drosophila* and humans[45–47]. Of course, the LTEE may differ from many natural populations in important respects including its low mutation rate, the absence of sex or horizontal gene transfer, and a stable environment. As we showed, high mutation rates tend to obscure the role of selection in molecular evolution. The effects of horizontal gene transfer[48] and variable environments[49,50] on the dynamic coupling of genomic and adaptive evolution should also be examined further. Long-term experiments with microorganisms provide opportunities for rigorous analyses of these issues.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 January; accepted 23 June 2016.
Published online 1 August 2016.

1. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74,** 1111–1120 (2004).
2. Hufford, M. B. *et al.* Comparative population genomics of maize domestication and improvement. *Nature Genet.* **44,** 808–811 (2012).
3. Vonholdt, B. M. *et al.* Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature* **464,** 898–902 (2010).
4. Lieberman, T. D. *et al.* Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nature Genet.* **43,** 1275–1280 (2011).
5. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24,** 1586–1591 (2007).
6. Lynch, M. & Conery, J. S. The origins of genome complexity. *Science* **302,** 1401–1404 (2003).
7. Whitney, K. D. & Garland, T. Jr Did genetic drift drive increases in genome complexity? *PLoS Genet.* **6,** e1001080 (2010).
8. Wichman, H. A., Badgett, M. R., Scott, L. A., Boulianne, C. M. & Bull, J. J. Different trajectories of parallel evolution during viral adaptation. *Science* **285,** 422–424 (1999).
9. Barrick, J. E. *et al.* Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* **461,** 1243–1247 (2009).
10. Tenaillon, O. *et al.* The molecular diversity of adaptive convergence. *Science* **335,** 457–461 (2012).
11. Lang, G. I. *et al.* Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* **500,** 571–574 (2013).
12. Kvitek, D. J. & Sherlock, G. Whole genome, whole population sequencing reveals that loss of signaling networks is the major adaptive strategy in a constant environment. *PLoS Genet.* **9,** e1003972 (2013).
13. Burke, M. K. *et al.* Genome-wide analysis of a long-term evolution experiment with *Drosophila*. *Nature* **467,** 587–590 (2010).
14. Levy, S. F. *et al.* Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* **519,** 181–186 (2015).
15. Khan, A. I., Dinh, D. M., Schneider, D., Lenski, R. E. & Cooper, T. F. Negative epistasis between beneficial mutations in an evolving bacterial population. *Science* **332,** 1193–1196 (2011).
16. Lenski, R. E., Rose, M. R., Simpson, S. C. & Tadler, S. C. Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2000 generations. *Am. Nat.* **138,** 1315–1341 (1991).
17. Wiser, M. J., Ribeck, N. & Lenski, R. E. Long-term dynamics of adaptation in asexual populations. *Science* **342,** 1364–1367 (2013).
18. Sniegowski, P. D., Gerrish, P. J. & Lenski, R. E. Evolution of high mutation rates in experimental populations of *E. coli*. *Nature* **387,** 703–705 (1997).
19. Blount, Z. D., Barrick, J. E., Davidson, C. J. & Lenski, R. E. Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature* **489,** 513–518 (2012).
20. Wielgoss, S. *et al.* Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proc. Natl Acad. Sci. USA* **110,** 222–227 (2013).
21. Woods, R., Schneider, D., Winkworth, C. L., Riley, M. A. & Lenski, R. E. Tests of parallel molecular evolution in a long-term experiment with *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **103,** 9107–9112 (2006).
22. Rozen, D. E. & Lenski, R. E. Long-term experimental evolution in *Escherichia coli*. VIII. Dynamics of a balanced polymorphism. *Am. Nat.* **155,** 24–35 (2000).
23. Plucain, J. *et al.* Epistasis and allele specificity in the emergence of a stable polymorphism in *Escherichia coli*. *Science* **343,** 1366–1369 (2014).
24. Deatherage, D. E. & Barrick, J. E. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using *breseq*. *Methods Mol. Biol.* **1151,** 165–188 (2014).
25. Barrick, J. E. *et al.* Identifying structural variation in haploid microbial genomes from short-read resequencing data using *breseq*. *BMC Genomics* **15,** 1039 (2014).
26. Chao, L., Vargas, C., Spear, B. B. & Cox, E. C. Transposable elements as mutator genes in evolution. *Nature* **303,** 633–635 (1983).
27. Tenaillon, O., Taddei, F., Radman, M. & Matic, I. Second-order selection in bacterial evolution: selection acting on mutation and recombination rates in the course of adaptation. *Res. Microbiol.* **152,** 11–16 (2001).
28. Gerrish, P. J. & Lenski, R. E. The fate of competing beneficial mutations in an asexual population. *Genetica* **102–103,** 127–144 (1998).
29. Barrick, J. E. & Lenski, R. E. Genome dynamics during experimental evolution. *Nature Rev. Genet.* **14,** 827–839 (2013).
30. Good, B. H. & Desai, M. M. Deleterious passengers in adapting populations. *Genetics* **198,** 1183–1208 (2014).
31. Maddamsetti, R., Lenski, R. E. & Barrick, J. E. Adaptation, clonal interference, and frequency-dependent interactions in a long-term evolution experiment with *Escherichia coli*. *Genetics* **200,** 619–631 (2015).
32. Gillespie, J. H. Genetic drift in an infinite population. The pseudohitchhiking model. *Genetics* **155,** 909–919 (2000).

33. Neher, R. A. & Shraiman, B. I. Genetic draft and quasi-neutrality in large facultatively sexual populations. *Genetics* **188,** 975–996 (2011).
34. Kosheleva, K. & Desai, M. M. The dynamics of genetic draft in rapidly adapting populations. *Genetics* **195,** 1007–1025 (2013).
35. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, 1983).
36. Sharp, P. M., Emery, L. R. & Zeng, K. Forces that influence the evolution of codon bias. *Phil. Trans. R. Soc. Lond. B* **365,** 1203–1212 (2010).
37. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nature Rev. Genet.* **12,** 32–42 (2011).
38. Stern, D. L. Evolutionary developmental biology and the problem of variation. *Evolution* **54,** 1079–1091 (2000).
39. Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134,** 25–36 (2008).
40. Oren, Y. *et al.* Transfer of noncoding DNA drives regulatory rewiring in bacteria. *Proc. Natl Acad. Sci. USA* **111,** 16112–16117 (2014).
41. Kibota, T. T. & Lynch, M. Estimate of the genomic mutation rate deleterious to overall fitness in *E. coli*. *Nature* **381,** 694–696 (1996).
42. Cooper, V. S., Schneider, D., Blot, M. & Lenski, R. E. Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli* B. *J. Bacteriol.* **183,** 2834–2841 (2001).
43. Miskinyte, M. *et al.* The genetic basis of *Escherichia coli* pathoadaptation to macrophages. *PLOS Pathog.* **9,** e1003802 (2013).
44. Wielgoss, S., Bergmiller, T., Bischofberger, A. M. & Hall, A. R. Adaptation to parasites and costs of parasite resistance in mutator and nonmutator bacteria. *Mol. Biol. Evol.* **33,** 770–782 (2016).
45. Charlesworth, J. & Eyre-Walker, A. The rate of adaptive evolution in enteric bacteria. *Mol. Biol. Evol.* **23,** 1348–1356 (2006).
46. Sawyer, S. A., Kulathinal, R. J., Bustamante, C. D. & Hartl, D. L. Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. *J. Mol. Evol.* **57** (suppl. 1), S154–S164 (2003).
47. Bustamante, C. D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437,** 1153–1157 (2005).
48. Cooper, T. F. Recombination speeds adaptation by reducing competition between beneficial mutations in populations of *Escherichia coli*. *PLoS Biol.* **5,** e225 (2007).
49. Satterwhite, R. S. & Cooper, T. F. Constraints on adaptation of *Escherichia coli* to mixed-resource environments increase over time. *Evolution* **69,** 2067–2078 (2015).
50. Paterson, S. *et al.* Antagonistic coevolution accelerates molecular evolution. *Nature* **464,** 275–278 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank N. Hajela for assistance, R. Maddamsetti and Z. Blount for discussions, and M. Lynch for starting the MAE lines. This research was supported by the US National Science Foundation (DEB-1451740 to R.E.L.), BEACON Center for the Study of Evolution in Action (DBI-0939454), European Research Council (FP7 grant 310944 to O.T.), European Union (FP7 grant 610427 to D.S.), French National Funding Agency (ANR-08-GENM-023-001 to D.S., O.T. and C.M.), French CNRS International Associated Laboratory (to D.S. and R.E.L.), and US National Institutes of Health (R00-GM087550 to J.E.B.). D.E.D. was supported by a traineeship from the Cancer Prevention and Research Institute of Texas. We acknowledge the use of high-performance computing resources at the Texas Advanced Computing Center.

Author Contributions O.T., J.E.B., D.S. and R.E.L. conceived the project; R.E.L. and J.L.B. provided strains; O.T., J.E.B., D.E.D., A.D., G.C.W., S.W., S.C. and C.M. analysed genomes and generated other data; N.R. developed theory; R.E.L., O.T. and J.E.B. wrote the paper. All authors approved the submitted version.

Author Information All sequencing data sets are available in the NCBI BioProject database under accession number PRJNA294072. The *breseq* analysis pipeline is available at GitHub (http://github.com/barricklab/breseq). Other analysis scripts are available at the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.6226d). R.E.L. will make strains available to qualified recipients, subject to a material transfer agreement. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.E.L. (lenski@msu.edu).

Reviewer Information *Nature* thanks M. Desai, G. Sherlock and C. Zeyl for their contribution to the peer review of this work.

METHODS

Long-term evolution experiment. The LTEE has 12 populations founded from two almost identical strains of *Escherichia coli*. Six populations, designated Ara−1 to Ara−6, started from REL606, a descendant of the B strain of Luria and Delbrück[51–53]. The other six, Ara+1 to Ara+6, derive from REL607, which differs from REL606 by point mutations in *araA* and *recD*. The mutation in *araA* was selected before starting the LTEE; it confers the ability to grow on L-arabinose, which provides a marker in competition assays used to measure fitness[16,17]. The *recD* mutation arose inadvertently before starting the LTEE. The LTEE began in 1988, and the populations have been propagated (with occasional interruptions) at 37 °C by daily 100-fold dilutions in 10 ml Davis minimal medium with 25 μg/ml glucose (http://lenski.mmg.msu.edu/ecoli/dm25liquid.html). The regrowth allows ∼6.67 generations per day; the population size fluctuates between $\sim 3 \times 10^6$ and $\sim 3 \times 10^8$ cells except in population Ara−3, which has had a population size several times larger since ∼33,000 generations, when cells gained the ability to consume the citrate that is also present in the medium[19,54]. Whole-population samples are taken every 75th transfer (500 generations) and stored with glycerol as a cryoprotectant at −80 °C, where they are available for later analysis. Here we analysed the genomes of two clones sampled from each population at 500, 1,000, 1,500, 2,000, 5,000, 10,000, 15,000, 20,000, 30,000, 40,000 and 50,000 generations (Supplementary Data 1). We deliberately included clones from the deeply diverged lineages in population Ara−2 from 20,000 generations onwards and both the majority Cit^+ lineage and the minority Cit^- lineage in population Ara−3 at generation 40,000. This sampling scheme does not affect inferences about the rates and patterns of genome evolution because both populations were hypermutable at these time points and thus excluded from the main analyses. These clones were included to illustrate diversity within populations, although we also found previously unknown cases of divergent lineages. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded during experiments and outcome assessment.

Mutation-accumulation experiment. The 15 MAE lines analysed here started from strain REL1207, which is an Ara^+ mutant of a clone sampled from LTEE population Ara−1 at 2,000 generations. REL1207 differs from REL606 by a total of eight mutations, including one in *araA* that confers the Ara^+ marker phenotype. Each line was propagated through 550 single-cell bottlenecks by picking a colony at random from a Davis minimal agar plate with glucose at 200 μg/ml and streaking the cells onto a fresh plate. Given ∼25 cell doublings to produce a typical colony[41], the 550 cycles represent ∼13,750 generations. The bottlenecks imposed by this procedure eliminate the genetic variation that fuels adaptation by natural selection; as a consequence, mutations accumulate at rates that depend on their underlying mutation rate but not their fitness effects, except for highly deleterious mutations that preclude sufficient growth to form a colony[29]. Because more mutations are deleterious than are beneficial, fitness declined under this regime (Extended Data Fig. 8). The 15 sequenced clonal isolates, each from a different MAE line, are JEB807–JEB821 (Supplementary Data 1). None of the lineages became hypermutable based on their mutational signatures and the absence of significant heterogeneity in the total mutations accumulated (see main text). However, the mean per-generation rate at which synonymous mutations arose was ∼3.5-fold higher in the MAE lines than in the five LTEE populations that remained non-mutators for all 50,000 generations (Supplementary Data 4; $t_s = 3.0755$, $P = 0.0065$). This difference may reflect the different conditions in liquid and agar media, including the glucose concentration and local cell density, which might affect the reactive oxygen species that cells experience. The comparisons between the LTEE and MAE (Extended Data Fig. 9) would change if the underlying rates of the various types of mutation responded disproportionately to the different conditions in the MAE. That possibility seems implausible for the different classes of point mutation (Extended Data Fig. 9a, b), and the differences would have to be substantially larger than the different rates of synonymous mutations to produce the excess IS*150* insertions (Extended Data Fig. 9c) and large indels (Extended Data Fig. 9f) observed in the LTEE relative to the MAE.

Genome sequencing. Frozen samples from the LTEE and MAE were revived via overnight growth at 37 °C in either LB or Davis minimal medium supplemented with 1,000 μg/ml glucose. Genomic DNA was isolated from each culture using the Qiagen Genomic-tip 100/G kit or equivalent. The DNA samples were sequenced at Genoscope or Integragen SA (Évry, France), the Michigan State University Research Technology Support Facility (East Lansing, USA), or the University of Texas at Austin Genome Sequencing and Analysis Facility (Austin, USA). Illumina Genome Analyzer and HiSeq instruments were used to generate single-end or paired-end reads ranging in length from 35 to 150 bases according to standard procedures, with median coverage of 80-fold and 95-fold for the 264 LTEE and 15 MAE clones, respectively (Supplementary Data 1). Of the 264 LTEE genomes in this study, 40 were previously analysed in other studies[9,19,20,55–57]. Supplementary Data 4 shows the number of every type of mutation inferred after performing the analyses described below on each of the LTEE and MAE genomes used in this study.

Mutation calling. We used *breseq* (versions 0.26.0 to 0.27.0) to predict both single-nucleotide and structural differences[24,25] based on how the Illumina reads for each sample mapped to the genome sequence of *E. coli* B REL606 (GenBank accession NC_012967.1)[52]. We counted and classified mutations using an updated version of the REL606 reference genome with improved feature annotations. The updated genome file (in both GenBank and GFF3 formats) and lists of predicted mutations in each evolved genome (in the Genome Diff format described in an appendix to the *breseq* manual) are freely available online (http://github.com/barricklab/LTEE-Ecoli).

Most types of single-step mutations, including large deletions and transposition events leading to copies of IS elements at new positions in the genome, are directly predicted by *breseq* when they occur in non-repetitive genomic regions. The initial lists of predicted mutations were curated and refined as previously described[24]. Briefly, complex mutations involving multiple steps (such as a new IS insertion followed by a flanking deletion) and structural mutations that overlap repetitive regions of the genome were manually resolved from unassigned new junction and missing coverage evidence in the *breseq* output. Large duplications and amplifications were detected by examining the coverage depth of mapped reads across the reference genome and comparing this information with the positions of repeat sequences and unassigned junctions. Owing to limitations of short-read DNA sequencing data, we could not fully predict point mutations and indels of one to a few base pairs within repeat regions (for example, IS elements) or gene conversions, in which intragenomic recombination between nearly identical copies of a large repeat region (for example, the seven copies of the rRNA operon) converts a minor variation in one copy to match exactly the sequence of another copy. Instead, all such genetic changes in repetitive regions of the genome were uniformly ignored in downstream analyses, as described later.

To validate the final lists of mutations predicted in each clone, we applied these changes to the ancestral REL606 sequence and used *breseq* to compare the Illumina reads against this simulated evolved genome to verify there were no further, unexplained discrepancies. This step of applying mutations to the reference genome was also used to estimate the final genome size of each evolved clone, with the assumption that new IS insertions were of the most common size for that IS element in the reference genome.

For 6 of the 264 LTEE samples, there was evidence of non-clonality in the sequence data. Some samples appeared to be mixtures of two very closely related clones that shared nearly all mutations but had one to several mutations specific to each type, together adding to a frequency of 100% (for example, sets of mutations at frequencies of 35% and 65%). This situation might result from inadvertently sampling two adjacent colonies on an agar plate when picking clones from an LTEE population. In other cases, only one or two mutations were found at an intermediate frequency. This type of heterogeneity might arise from strong selection favouring new mutations during colony outgrowth, subculturing and revival of samples before DNA extraction, as these conditions differ from the LTEE. In each case, we reconstructed the major genotype in the sample, as noted in Supplementary Data 1.

We also ignored putative genome variation associated with a cryptic 186-like prophage element (REL606 genome coordinates 880528–904682). In ten of the LTEE populations, we observed clones with increased read-coverage depth of this region and reads spanning a new sequence junction consistent with either tandem head-to-tail amplifications of this region or the production of circular DNA molecules joined at these exact nucleotides. The changes in the apparent copy number of this region often deviated from the integer values expected for a stable duplication or amplification. The prophage-related changes in coverage appeared most often in genomes isolated from 2,000 generations or earlier in the LTEE. There is no evidence of infective phage production in the LTEE, but it is possible that replication of DNA encoding a defective phage occurs stochastically at some low level in the ancestral strain REL606 or that production of this DNA is induced by stress when culturing samples for DNA isolation.

Phylogenetic consistency. Owing to the long duration of the LTEE and the evolution of mutators in several lineages, some mutations may be hidden or initially grouped with other mutations into a single change when comparing a late-generation evolved genome with the ancestral genome. For example, a point mutation might occur early in the experiment and then the region containing that mutation is later deleted. Similarly, the deletion of one base early and the subsequent deletion of an adjacent base would be called as a single two-base deletion in later samples. To obtain more accurate counts in light of these issues, we used each population's inferred phylogeny to split or add mutations, as appropriate, so that the mutation list for each clone reflects the most parsimonious set of mutational steps between that clone and its ancestor. Specifically, we chose histories with the fewest total mutations, the fewest mutations on early branches (in case of ties), and the fewest

total nucleotide changes summed over all mutations. Because this procedure is conservative in adding mutations to achieve phylogenetic consistency, it might underestimate the number of mutations on branches leading to an evolved genome when intermediate states are not resolved by the relationships of the sequenced clones.

Final mutation lists. We performed two final filtering steps to enable the sets of mutations to be uniformly compared across all genomes. In doing so, we classified as 'small mutations' all single-nucleotide substitutions, insertions and deletions of 20 or fewer bp, substitutions replacing 20 or fewer bp in the reference genome with 20 or fewer other bp, and all simple sequence repeat (SSR) mutations regardless of their size. SSR mutations add or remove one or more copies of a tandem-repeat unit consisting of one or a few bp. We defined SSR mutations as containing at least two copies of the repeat unit and having a total length of at least five bp when including all copies of the tandem repeat in the reference genome. For example, the genetic changes GGGGG→GGGG, TATATA→TATATATA and TACGTTACGT→TACGT would all be classified as SSR mutations, but GGGG→GGGGG, TATA→TATATA and TACGT→TACGTTACGT would not. All other genomic changes were considered 'large mutations' for purposes of filtering.

The ability to call small mutations located in repetitive regions of the genome is dependent on read length, so we removed all such mutations in regions where it would be a problem to uniformly detect them from the mutation lists before further analyses. To do this, we enumerated all regions of ≥20 bp that had an exact match elsewhere in the genome of the ancestral strain REL606 using MUMmer v.3.23 (ref. 58). We then merged regions from this list that were separated by five or fewer bp. All resulting regions that were now ≥35 bp were included in a list of masked genomic intervals. We also added to this list a hypervariable SSR consisting of seven copies of a tetranucleotide sequence that could not be reliably called in data sets with short reads (coordinates 2103889–2103919). Any small mutations contained in these masked regions were excluded from all downstream analyses.

Finally, we flagged all nucleotide substitutions or small indels occurring within 20 bp of the end of an IS element. The sequences directly adjacent to IS elements appear to experience an unusually high mutation rate, possibly due to frequent transposase cleavage and DNA repair. Mutations at these IS-adjacent sites probably have no effect on cellular phenotypes and fitness. We excluded them from the final lists of mutations used in all further analyses because they could bias the inferred mutational spectra and rates.

Phylogenetic analyses. To produce the phylogenetic trees shown in Fig. 2, we used the point mutations associated with each clone. A minimum-evolution tree was built using the Jukes–Cantor one-parameter model[59]. We used this model for two reasons. First, the mutator lineages had very different mutational spectra from the non-mutators[9,20,55,57]. Second, many mutations seen in non-mutator lineages were under positive selection, and so it is appropriate to give the mutations equal weight and not, for instance, reduce the importance of transitions relative to transversions. The trees were plotted with the R package APE[60]. The composite tree has the star-like structure expected for independent evolution of the populations. Therefore, trees were made separately for each population and then combined in Fig. 2, which allowed multiple basal branches to be placed with the appropriate populations.

Parallel evolution in non-mutator lineages. For genomes that did not come from point-mutation hypermutator lineages (Supplementary Data 1), we examined the extent of parallelism at the gene level in two ways. The first approach was based only on nonsynonymous mutations, because it is straightforward to quantify the overall extent of parallelism, determine the statistical significance of the parallelism, and rank genes based on their contributions to the significance. For each protein-coding gene i, we know its length, L_i, and the number of independent nonsynonymous mutations observed in that gene across all clones from non-mutator and premutator lineages, N_i. We summed the lengths and relevant mutations over all single-copy protein-coding genes in the ancestral genome to obtain L_{tot} (3,920,306) and N_{tot} (457, including two mutations that each affected overlapping reading frames), respectively. We computed the expected number of mutations in each gene, E_i, as follows:

$$E_i = N_{tot} (L_i / L_{tot})$$

We then computed a G_i score for each gene for which $N_i > 0$ as follows:

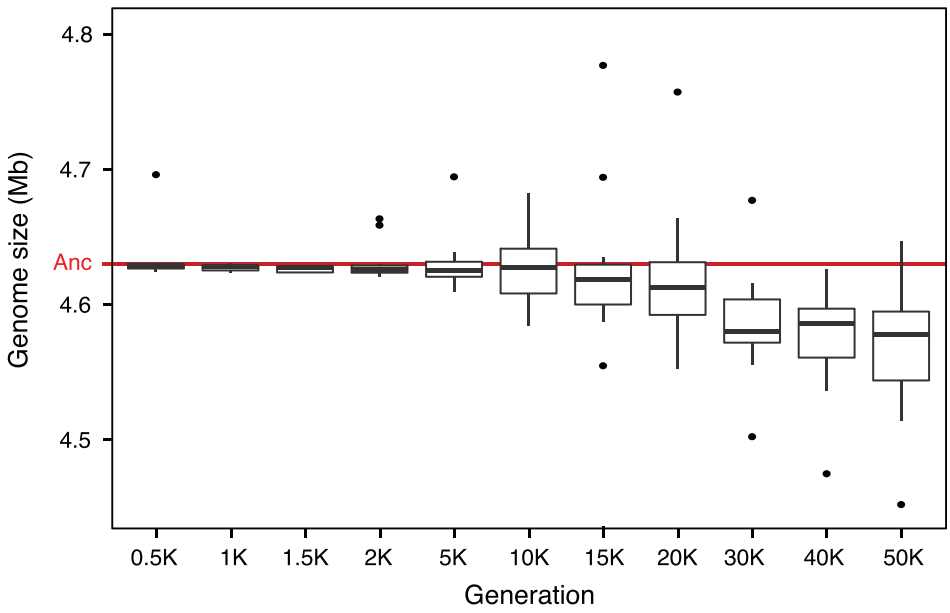
$$G_i = 2N_i \log_e(N_i / E_i)$$

We set $G_i = 0$ for those genes for which $N_i = 0$. This analysis ignores variability among genes in the proportion of sites at risk for nonsynonymous mutations. However, such differences are small and should hardly affect the analysis. The total G statistic equals the sum of the scores over all genes. To compute the expected G statistic under the null hypothesis of a random distribution of mutations, we generated 1,000 simulated data sets in which N_{tot} mutations were randomly placed throughout the coding genome. We computed the total G statistic for each simulated data set, and we calculated its mean and standard deviation across the 1,000 simulations. To assess the significance of the observed G statistic, we computed the Z score as the difference between the observed and mean simulated values, divided by the standard deviation of the simulated values. Supplementary Data 2 lists each gene and the information used to calculate its G score. Table 1 shows the 15 genes with the highest G scores.

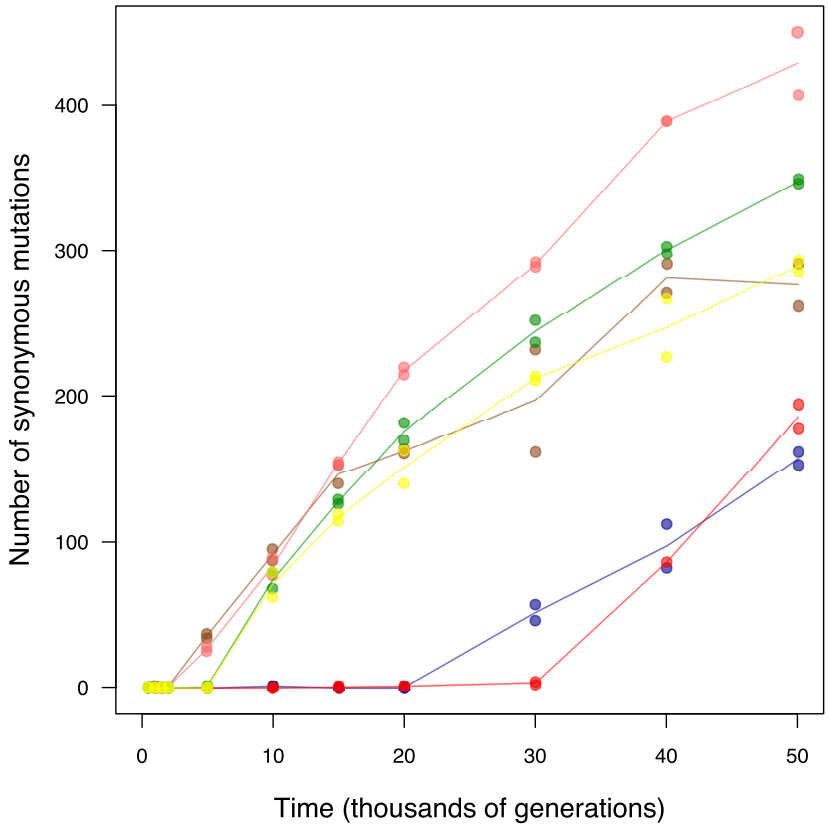
Supplementary Data 2 also shows other categories of mutation in or near each protein-coding gene including synonymous mutations, intergenic point mutations (between any particular gene and one of its immediately adjacent genes), IS insertions, small indels (≤50 bp), large deletions (>50 bp) and long duplications (>50 bp). Table 2 shows the 16 genes that had the most total deletions, duplications, insertions and intergenic point mutations (that is, all mutations except synonymous and nonsynonymous mutations in the coding gene itself).

Parallel evolution in mutator lineages. We examined parallel changes in lineages that evolved point-mutation hypermutability by analysing nonsynonymous substitutions as above. To identify mutations that occurred after a lineage became hypermutable (Supplementary Data 3), we subtracted the mutations that occurred on non-mutator branches from the total mutations. This approach may result in a few mutations that arose before hypermutability being included in the counts for mutator lineages, but given the large increases in the point-mutation rate in the mutators (Fig. 1) it provides a reasonable approximation.

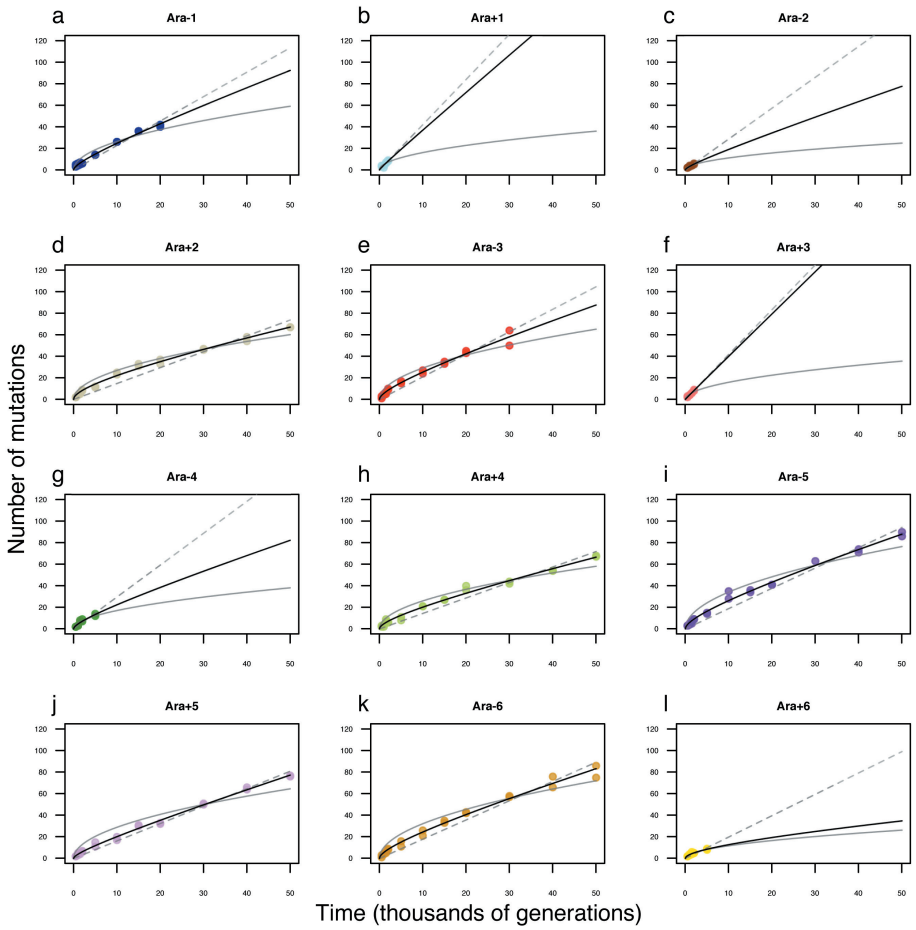
51. Daegelen, P., Studier, F. W., Lenski, R. E., Cure, S. & Kim, J. F. Tracing ancestors and relatives of *Escherichia coli* B, and the derivation of B strains REL606 and BL21(DE3). *J. Mol. Biol.* **394,** 634–643 (2009).
52. Jeong, H. *et al.* Genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3). *J. Mol. Biol.* **394,** 644–652 (2009).
53. Luria, S. E. & Delbrück, M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28,** 491–511 (1943).
54. Blount, Z. D., Borland, C. Z. & Lenski, R. E. Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **105,** 7899–7906 (2008).
55. Wielgoss, S. *et al.* Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. *G3 (Bethesda)* **1,** 183–186 (2011).
56. Raeside, C. *et al.* Large chromosomal rearrangements during a long-term evolution experiment with *Escherichia coli*. *mBio* **5,** e01377–14 (2014).
57. Maddamsetti, R. *et al.* Synonymous genetic variation in natural isolates of *Escherichia coli* does not predict where synonymous substitutions occur in a long-term experiment. *Mol. Biol. Evol.* **32,** 2897–2904 (2015).
58. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5,** R12 (2004).
59. Desper, R. & Gascuel, O. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comput. Biol.* **9,** 687–705 (2002).
60. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20,** 289–290 (2004).



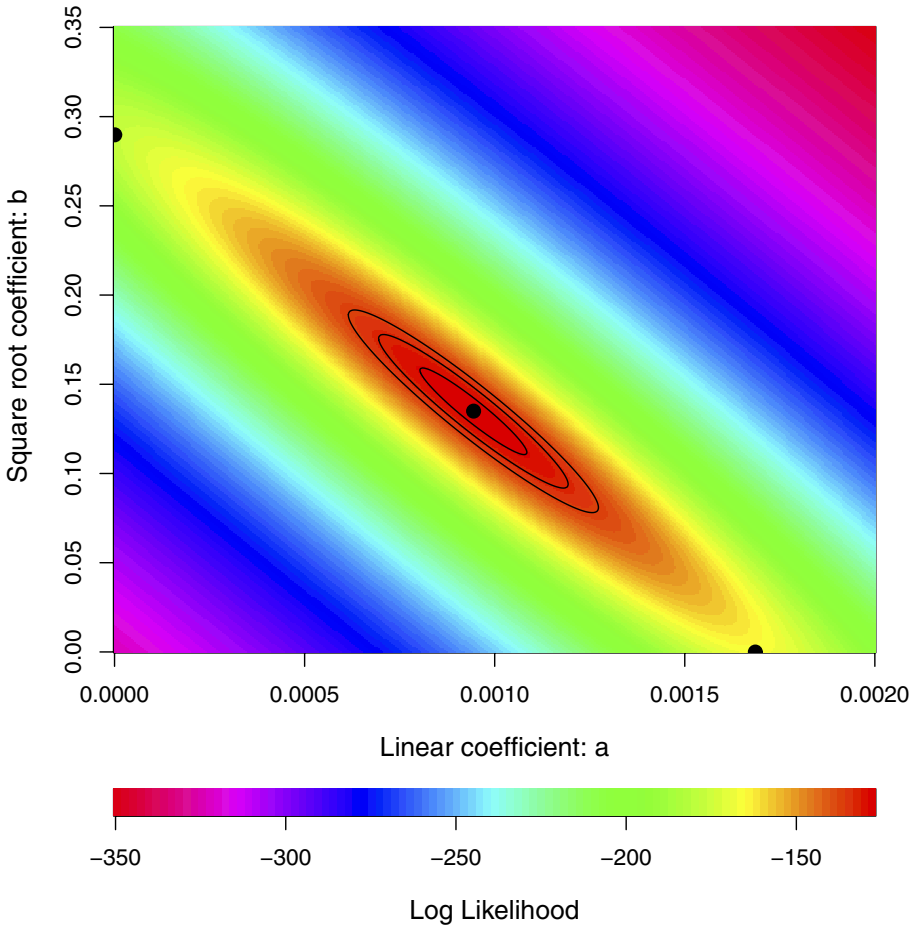
Extended Data Figure 1 | Changes in genome size during the LTEE. Box-and-whiskers plot showing the distribution of average genome length (Mb) for each of the 12 LTEE populations based on the two clones sequenced at each time point shown from 500 to 50,000 generations. The red line shows the length of the ancestral genome. The boxes are the interquartile range (IQR), which spans the second and third quartiles of the data (25th to 75th percentiles); the thick black lines are medians; the whiskers extend to the outermost values that are within 1.5 times the IQR; and the points show all outlier values beyond the whiskers.



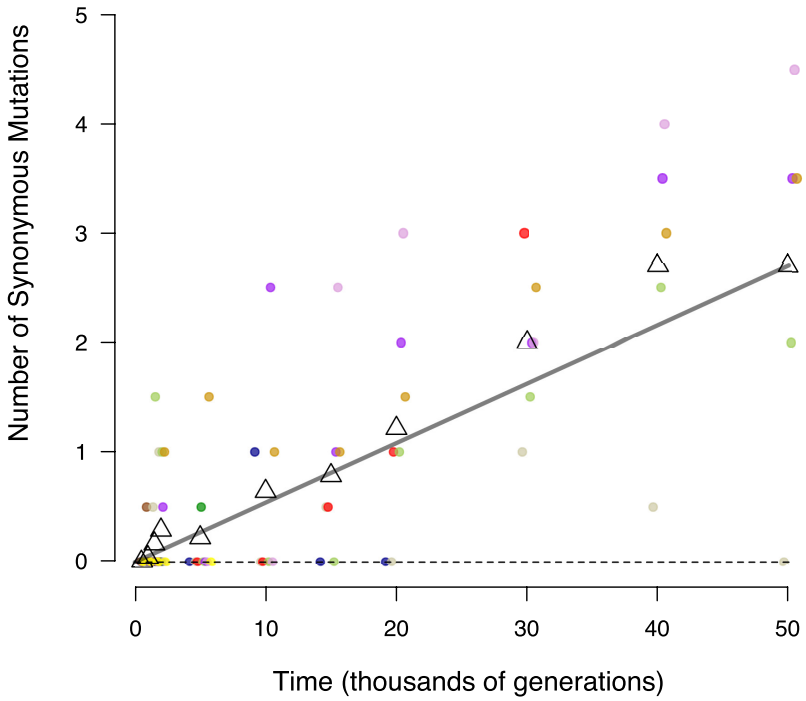
Extended Data Figure 2 | Accumulation of synonymous mutations in populations that evolved point-mutation hypermutability. Each symbol shows a sequenced genome from a hypermutable lineage. Colours are the same as those in Fig. 1. The accumulation of synonymous substitutions serves as a proxy for the underlying point-mutation rate. All four of the populations that became hypermutable before 10,000 generations accumulated synonymous mutations at higher rates between 10,000 and 20,000 generations than between 40,000 and 50,000 generations, indicating the evolution of reduced mutability.



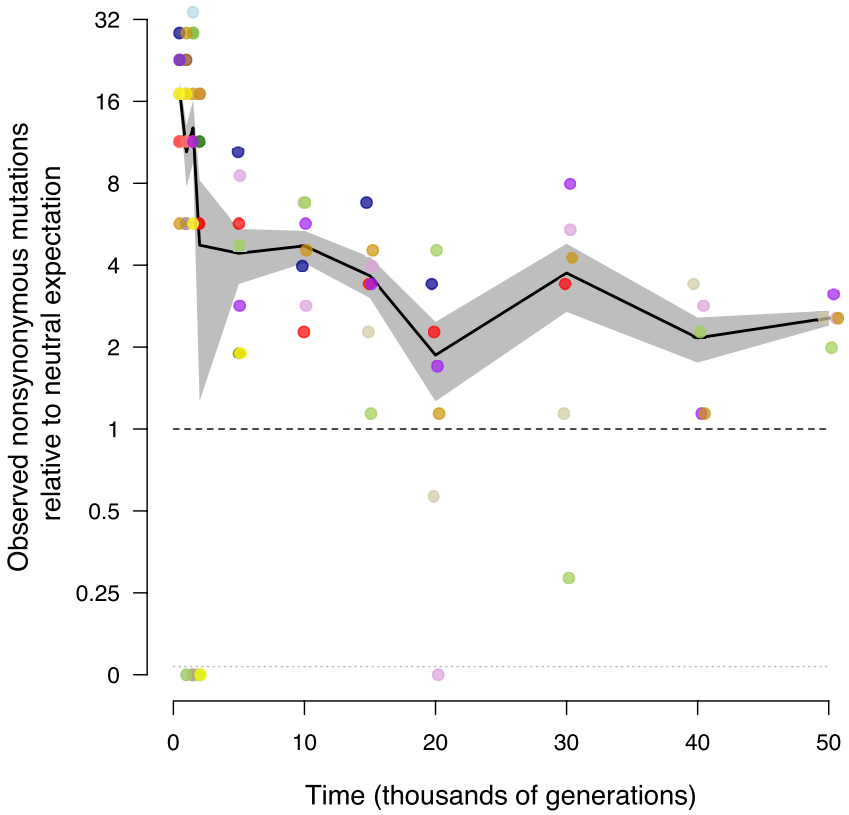
Extended Data Figure 3 | Alternative models fit to trajectory of genome evolution for each LTEE population. a, Ara−1. **b**, Ara+1. **c**, Ara−2. **d**, Ara+2. **e**, Ara−3. **f**, Ara+3. **g**, Ara−4. **h**, Ara+4. **i**, Ara−5. **j**, Ara+5. **k**, Ara−6. **l**, Ara+6. Each symbol shows the total mutations in a sequenced genome; in many cases, the symbols for the two genomes from the same population and generation are not distinguishable because they have the same, or almost the same, number of mutations. For the populations that evolved hypermutability, data are shown only for time points before mutators arose. In each panel, the dashed grey line shows the best fit to the linear model; the solid grey curve shows the best fit to the square-root model; and the solid black curve shows the best fit to the composite model with both linear and square-root terms.



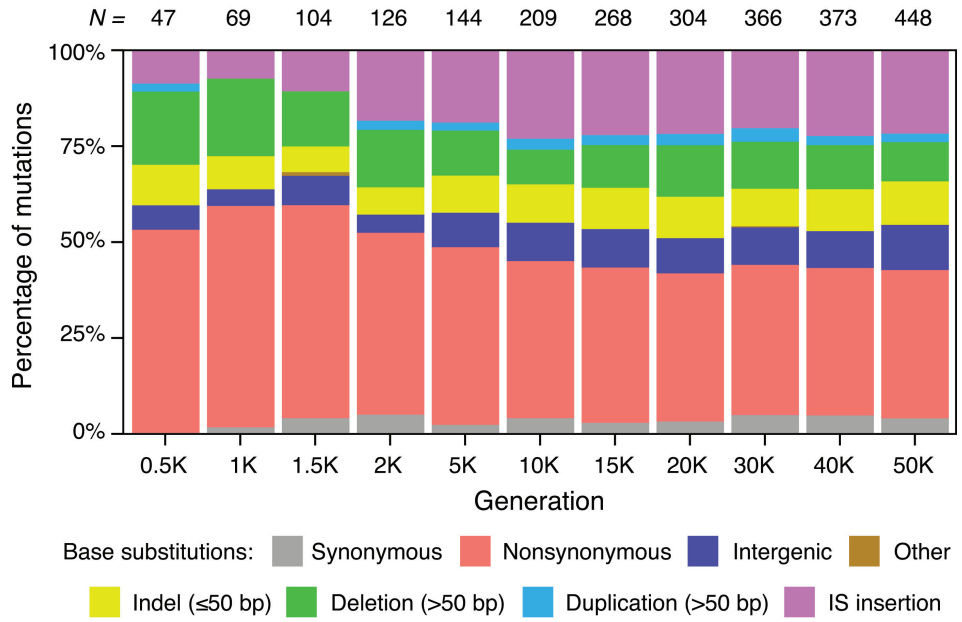
Extended Data Figure 4 | Uncertainty in parameter estimation for the model describing the rates of accumulation for neutral and beneficial mutations. Contours show relative likelihoods for simultaneously estimating the linear and square-root coefficients from the observed numbers of mutations that accumulated over time in non-mutator and premutator lineages (Fig. 3). The black central point shows the maximum likelihood estimates, and the three black contours show solutions 2, 6 and 10 log units away. The points on the horizontal and vertical axes show values for the best one-parameter models.



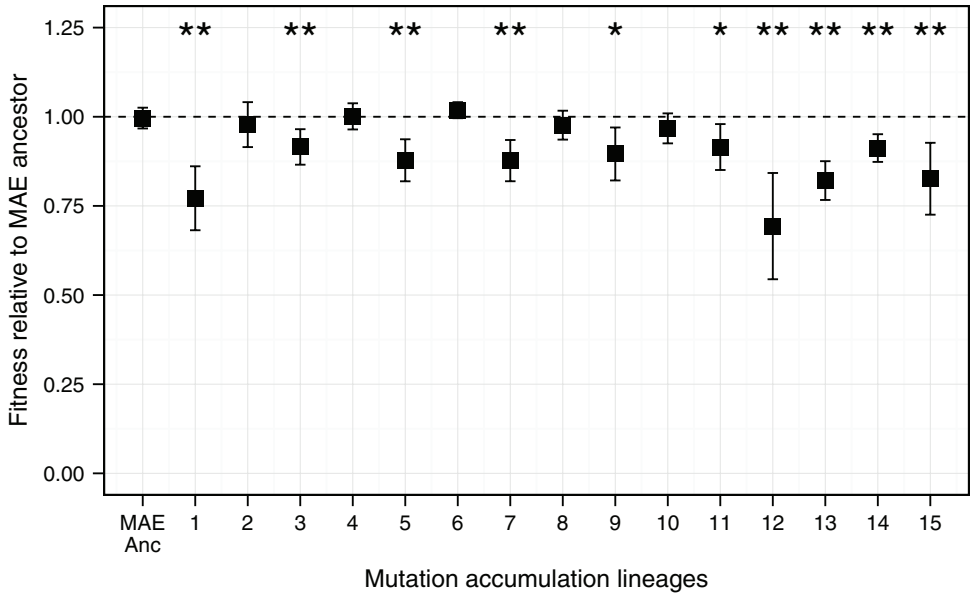
Extended Data Figure 5 | Accumulation of synonymous substitutions in non-mutator lineages. Each filled symbol shows the mean number of synonymous mutations in the (usually two) non-mutator genomes from an LTEE population that were sequenced at that time point; non-integer values can occur if the two genomes have different numbers. Small horizontal offsets were added so that overlapping points are visible. Colours are the same as in Fig. 1. Open triangles show the grand means of the replicate populations. The grey line extends from the intercept to the final grand mean. The slope of that line was used to scale the relative rates of synonymous, nonsynonymous and intergenic point mutations in Fig. 4.



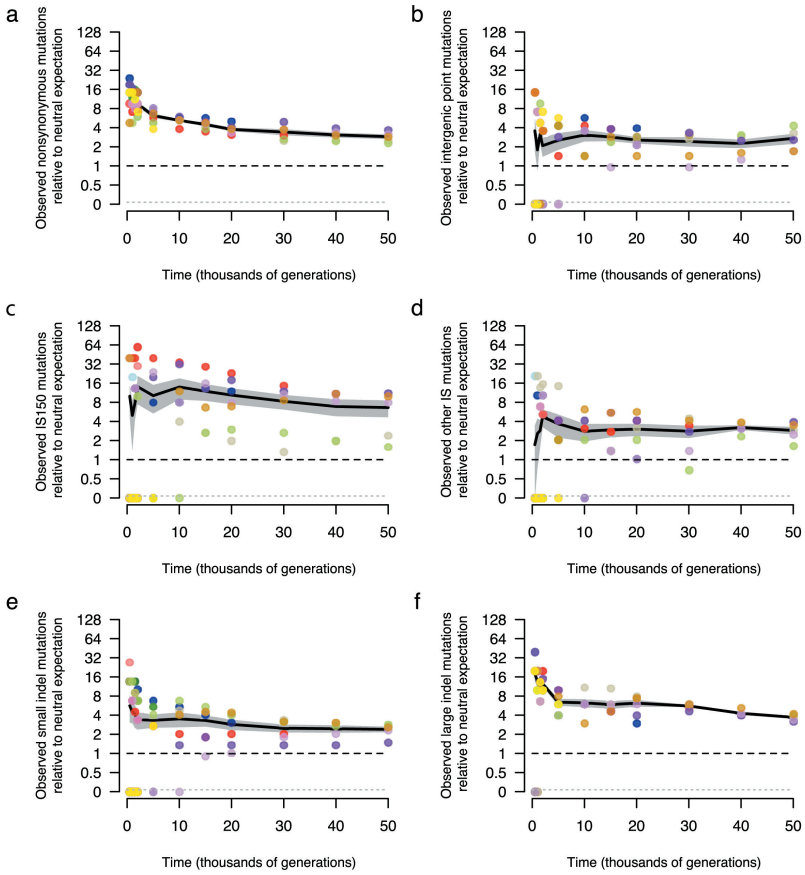
Extended Data Figure 6 | Temporal trend in accumulation of nonsynonymous mutations relative to the neutral expectation in non-mutator lineages. Interval-specific accumulation of nonsynonymous mutations calculated from changes in the total number of nonsynonymous mutations between successive samples. As with the cumulative data in Fig. 4b, values are scaled by the average rate of accumulation for synonymous mutations over 50,000 generations, after adjusting for the numbers of genomic sites at risk for nonsynonymous and synonymous mutations. Each point shows the average rate calculated for a non-mutator or premutator population; small horizontal offsets were added so that overlapping points are visible. Note the discontinuous scale; populations with no additional mutations over an interval are plotted below. Colours are the same as in Fig. 1. Black lines connect grand means; the grey shading shows standard errors calculated from the replicate populations.



Extended Data Figure 7 | Mutational spectrum for non-mutator lineages in the LTEE. Shaded bars show the distribution of different types of genetic change for all independent mutations found in the set of non-mutator clones that were sequenced at each generation. The total number of mutations in this set at each time point (*N*) is shown above each column. Base substitutions are divided into synonymous, nonsynonymous, intergenic, and other categories; the nonsynonymous category includes nonsense mutations, and the 'other' category includes rare point mutations in noncoding RNA genes and pseudogenes.



Extended Data Figure 8 | Changes in fitness of MAE lines after 550 single-cell bottlenecks and ~13,750 generations. Each point shows the mean fitness based on nine competition assays between the MAE ancestor (REL1207) or one of the 15 MAE lineages (JEB807–JEB821) and the Ara$^-$ variant of the MAE ancestor (REL1206). One-day competition assays were performed using the standard procedures and same conditions as for the LTEE[16,17]. Error bars show 95% confidence intervals. $*P < 0.05$, $**P < 0.01$, based on two-tailed t-tests of the null hypothesis that relative fitness equals 1. Ten of the fifteen MAE lines experienced significant fitness declines, while none had significant gains.



Extended Data Figure 9 | Trajectories for mutations by class in the LTEE in comparison with neutral expectations based on the MAE. **a**–**f**, Accumulation of nonsynonymous mutations (**a**), intergenic point mutations (**b**), IS*150* insertions (**c**), all other IS-element insertions (**d**), small indels (**e**) and large indels (**f**). Colours are the same as in Fig. 1. All values are expressed relative to the rate at which synonymous mutations accumulated in non-mutator LTEE lineages over 50,000 generations (Fig. 4a), and then scaled by the ratio of the number of the indicated class of mutation relative to the number of synonymous mutations in the MAE lines. In all panels, each symbol shows a non-mutator or premutator population. Note the discontinuous scale, in which populations with no mutations of the indicated type are plotted below. Black lines connect grand means over the replicate LTEE populations; the grey shading shows the corresponding standard errors.

Bibliography

[1] Eltaf Alamyar, Véronique Giudicelli, Shuo Li, Patrice Duroux, and Marie Paule Lefranc. IMGT/Highv-quest: The IMGT web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. *Immunome Research*, 8(1), 2012.

[2] Ian J Amanna, Nichole E Carlson, and Mark K Slifka. Duration of humoral immunity to common viral and vaccine antigens. *The New England journal of medicine*, 357(19):1903–15, November 2007.

[3] Ian J Amanna and Mark K Slifka. Contributions of humoral and cellular immunity to vaccine-induced protection in humans. *Virology*, 411(2):206–15, March 2011.

[4] Ramy Arnaout, William Lee, Patrick Cahill, Tracey Honan, Todd Sparrow, Michael Weiand, Chad Nusbaum, Klaus Rajewsky, and Sergei B Koralov. High-resolution description of antibody heavy-chain repertoires in humans. *PloS one*, 6(8):e22365, January 2011.

[5] Jennifer Benichou, Rotem Ben-Hamo, Yoram Louzoun, and Sol Efroni. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology*, 135(3):183–91, March 2012.

[6] R Benner, W Hijmans, and J J Haaijman. The bone marrow: the major source of serum immunoglobulins, but still a neglected site of antibody formation. *Clinical and experimental immunology*, 46(1):1–8, October 1981.

[7] Nadia L Bernasconi, Elisabetta Traggiai, and Antonio Lanzavecchia. Maintenance of serological memory by polyclonal activation of human memory B cells. *Science (New York, N.Y.)*, 298(5601):2199–202, December 2002.

[8] Erin Bromage, Rebecca Stephens, and Lama Hassoun. The third dimension of ELISPOTs: quantifying antibody secretion from individual plasma cells. *Journal of immunological methods*, 346(1-2):75–9, July 2009.

[9] D R Burton, C F Barbas, M A Persson, S Koenig, R M Chanock, and R A Lerner. A large array of human monoclonal antibodies to type 1 human immunodeficiency virus from combinatorial libraries of asymptomatic seropositive individuals. *Proceedings of the National Academy of Sciences of the United States of America*, 88(22):10134–7, November 1991.

[10] G Cambridge, M J Leandro, M Teodorescu, J Manson, A Rahman, D A Isenberg, and J C Edwards. B cell depletion therapy in systemic lupus erythematosus: effect on autoantibody and antimicrobial antibody profiles. *Arthritis and rheumatism*, 54(11):3612–22, November 2006.

[11] Geraldine Cambridge, Maria J Leandro, Jonathan C W Edwards, Michael R Ehrenstein, Martin Salden, Mark Bodman-Smith, and Anthony D B Webster. Serologic changes following B lymphocyte depletion therapy for rheumatoid arthritis. *Arthritis and rheumatism*, 48(8):2146–54, August 2003.

[12] Zhiliang Chen, Andrew M Collins, Yan Wang, and Bruno A Gaëta. Clustering-based identification of clonally-related immunoglobulin gene sequence sets. *Immunome research*, 6 Suppl 1(Suppl 1):S4, September 2010.

[13] Graeme Cowan, Nicola J Weston-Bell, Dean Bryant, Anja Seckinger, Dirk Hose, Niklas Zojer, and Surinder S Sahota. Massive parallel IGHV gene sequencing reveals a germinal center pathway in origins of human multiple myeloma. *Oncotarget*, 6(15):13229–40, May 2015.

[14] Brandon J DeKosky, Gregory C Ippolito, Ryan P Deschner, Jason J Lavinder, Yariv Wine, Brandon M Rawlings, Navin Varadarajan, Claudia Giesecke, Thomas Dörner, Sarah F Andrews, Patrick C Wilson, Scott P Hunicke-Smith, C Grant Willson, Andrew D Ellington, and George Georgiou. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nature biotechnology*, 31(2):166–9, February 2013.

[15] Brandon J DeKosky, Takaaki Kojima, Alexa Rodin, Wissam Charab, Gregory C Ippolito, Andrew D Ellington, and George Georgiou. In-depth

determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nature Medicine*, 21(1):86–91, December 2014.

[16] Brandon J DeKosky, Oana I Lungu, Daechan Park, Erik L Johnson, Wissam Charab, Constantine Chrysostomou, Daisuke Kuroda, Andrew D Ellington, Gregory C Ippolito, Jeffrey J Gray, and George Georgiou. Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proceedings of the National Academy of Sciences of the United States of America*, 113(19):E2636–45, May 2016.

[17] Jonathan Desponds, Thierry Mora, and Aleksandra M Walczak. Fluctuating fitness shapes the clone-size distribution of immune repertoires. *Proceedings of the National Academy of Sciences of the United States of America*, 113(2):274–9, January 2016.

[18] Jared W Ellefson, Jimmy Gollihar, Raghav Shroff, Haridha Shivram, Vishwanath R Iyer, and Andrew D Ellington. Synthetic evolutionary origin of a proofreading reverse transcriptase. *Science (New York, N.Y.)*, 352(6293):1590–3, June 2016.

[19] D T Fearon, P Manders, and S D Wagner. Arrested differentiation, the self-renewing memory lymphocyte, and vaccination. *Science (New York, N.Y.)*, 293(5528):248–50, July 2001.

[20] Jessica A Finn and James E Crowe. Impact of new sequencing tech-

nologies on studies of the human B cell repertoire. *Current opinion in immunology*, 25(5):613–8, October 2013.

[21] Juan Flores-Montero, Ruth de Tute, Bruno Paiva, José Juan Perez, Sebastian Böttcher, Henk Wind, Luzalba Sanoja, Noemí Puig, Quentin Lecrevisse, María Belén Vidriales, Jacques J M van Dongen, and Alberto Orfao. Immunophenotype of normal vs. myeloma plasma cells: Toward antibody panel specifications for MRD detection in multiple myeloma. *Cytometry. Part B, Clinical cytometry*, June 2015.

[22] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, December 2012.

[23] George Georgiou, Gregory C Ippolito, John Beausang, Christian E Busse, Hedda Wardemann, and Stephen R Quake. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature biotechnology*, 32(2):158–68, February 2014.

[24] Jacob Glanville, Tracy C Kuo, H-Christian von Büdingen, Lin Guey, Jan Berka, Purnima D Sundar, Gabriella Huerta, Gautam R Mehta, Jorge R Oksenberg, Stephen L Hauser, David R Cox, Arvind Rajpal, and Jaume Pons. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proceedings of the National Academy of Sciences of the United States of America*, 108(50):20066–71, December 2011.

[25] Jacob Glanville, Wenwu Zhai, Jan Berka, Dilduz Telman, Gabriella Huerta, Gautam R Mehta, Irene Ni, Li Mei, Purnima D Sundar, Giles M R Day, David Cox, Arvind Rajpal, and Jaume Pons. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proceedings of the National Academy of Sciences of the United States of America*, 106(48):20216–21, December 2009.

[26] Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature reviews. Genetics*, 17(6):333–51, May 2016.

[27] Victor Greiff, Pooja Bhat, Skylar C Cook, Ulrike Menzel, Wenjing Kang, and Sai T Reddy. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome medicine*, 7(1):49, 2015.

[28] Jessica L. Halliley, Christopher M. Tipton, Jane Liesveld, Alexander F. Rosenberg, Jaime Darce, Ivan V. Gregoretti, Lana Popova, Denise Kaminiski, Christopher F. Fucile, Igor Albizua, Shuya Kyu, Kuang-Yueh Chiang, Kyle T. Bradley, Richard Burack, Mark Slifka, Erika Hammarlund, Hao Wu, Liping Zhao, Edward E. Walsh, Ann R. Falsey, Troy D. Randall, Wan Cheung Cheung, Iñaki Sanz, and F. Eun-Hyung Lee. Long-Lived Plasma Cells Are Contained within the CD19CD38hiCD138+ Subset in Human Bone Marrow. *Immunity*, 43(1):132–45, July 2015.

[29] Carole J Henry Dunand and Patrick C Wilson. Restricted, canonical, stereotyped and convergent immunoglobulin responses. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 370(1676), September 2015.

[30] Weichun Huang, Leping Li, Jason R Myers, and Gabor T Marth. ART: a next-generation sequencing read simulator. *Bioinformatics (Oxford, England)*, 28(4):593–4, February 2012.

[31] Gregory C Ippolito, Kam Hon Hoi, Sai T Reddy, Sean M Carroll, Xin Ge, Tobias Rogosch, Michael Zemlin, Leonard D Shultz, Andrew D Ellington, Carla L Vandenberg, and George Georgiou. Antibody repertoires in humanized NOD-scid-IL2Rγ(null) mice and human B cells reveals human-like diversification and tolerance checkpoints in the mouse. *PloS one*, 7(4):e35497, January 2012.

[32] Gregory C Ippolito, Robert L Schelonka, Michael Zemlin, Ivaylo I Ivanov, Ryoki Kobayashi, Cosima Zemlin, G Larry Gartland, Lars Nitschke, Jukka Pelkonen, Kohtaro Fujihashi, Klaus Rajewsky, and Harry W Schroeder. Forced usage of positively charged amino acids in immunoglobulin CDR-H3 impairs B cell development and antibody production. *The Journal of experimental medicine*, 203(6):1567–78, June 2006.

[33] Katherine J L Jackson, Yi Liu, Krishna M Roskin, Jacob Glanville, Ramona A Hoh, Katie Seo, Eleanor L Marshall, Thaddeus C Gurley, M Anthony Moody, Barton F Haynes, Emmanuel B Walter, Hua-Xin

Liao, Randy A Albrecht, Adolfo García-Sastre, Javier Chaparro-Riggers, Arvind Rajpal, Jaume Pons, Birgitte B Simen, Bozena Hanczaruk, Cornelia L Dekker, Jonathan Laserson, Daphne Koller, Mark M Davis, Andrew Z Fire, and Scott D Boyd. Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. *Cell host & microbe*, 16(1):105–14, July 2014.

[34] Denise A. Kaminski, Chungwen Wei, Yu Qian, Alexander F. Rosenberg, and Ignacio Sanz. Advances in Human B Cell Phenotypic Profiling. *Frontiers in Immunology*, 3:302, January 2012.

[35] Joseph Kaplinsky and Ramy Arnaout. Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples. *Nature communications*, 7:11881, June 2016.

[36] Jens C Krause, Tshidi Tsibane, Terrence M Tumpey, Chelsey J Huffman, Bryan S Briney, Scott A Smith, Christopher F Basler, and James E Crowe. Epitope-specific human influenza antibody repertoires diversify by B cell intraclonal sequence divergence and interclonal convergence. *Journal of immunology (Baltimore, Md. : 1950)*, 187(7):3704–11, October 2011.

[37] Martin Krzywinski, Jacqueline Schein, Inanç Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–45, September 2009.

[38] Kevin Larimore, Michael W McCormick, Harlan S Robins, and Philip D Greenberg. Shaping of human germline IgH repertoires revealed by deep sequencing. *Journal of immunology (Baltimore, Md. : 1950)*, 189(6):3221–30, September 2012.

[39] Uri Laserson, Francois Vigneault, Daniel Gadala-Maria, Gur Yaari, Mohamed Uduman, Jason A Vander Heiden, William Kelton, Sang Taek Jung, Yi Liu, Jonathan Laserson, Raj Chari, Je-Hyuk Lee, Ido Bachelet, Brendan Hickey, Erez Lieberman-Aiden, Bozena Hanczaruk, Birgitte B Simen, Michael Egholm, Daphne Koller, George Georgiou, Steven H Kleinstein, and George M Church. High-resolution antibody dynamics of vaccine-induced immune responses. *Proceedings of the National Academy of Sciences of the United States of America*, 111(13):4928–33, April 2014.

[40] Jason J Lavinder, Kam Hon Hoi, Sai T Reddy, Yariv Wine, and George Georgiou. Systematic characterization and comparative analysis of the rabbit immunoglobulin repertoire. *PloS one*, 9(6):e101322, January 2014.

[41] Jason J Lavinder, Andrew P Horton, George Georgiou, and Gregory C Ippolito. Next-generation sequencing and protein mass spectrometry for the comprehensive analysis of human cellular and serum antibody repertoires. *Current opinion in chemical biology*, 24:112–20, February 2015.

[42] T. Magoc and S. L. Salzberg. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21):2957–2963, November 2011.

[43] Anne E Magurran. Ecological diversity and its measurements. *Princeton University Press. New Jersey.*, page 177pp., 1988.

[44] R A Manz, A Thiel, and A Radbruch. Lifetime of plasma cells in the bone marrow. *Nature*, 388(6638):133–4, July 1997.

[45] Rudolf A Manz, Sergio Arce, Giuliana Cassese, Anja E Hauser, Falk Hiepe, and Andreas Radbruch. Humoral immunity and long-lived plasma cells. *Current opinion in immunology*, 14(4):517–21, August 2002.

[46] Pascale Mathonet and Christopher G Ullman. The Application of Next Generation Sequencing to the Understanding of Antibody Repertoires. *Frontiers in immunology*, 4:265, January 2013.

[47] R McMillan, R L Longmire, R Yelenosky, J E Lang, V Heath, and C G Craddock. Immunoglobulin synthesis by human lymphoid tissues: normal bone marrow as a major site of IgG production. *Journal of immunology (Baltimore, Md. : 1950)*, 109(6):1386–94, December 1972.

[48] Henrik E Mei, Ina Wirries, Daniela Frölich, Mikael Brisslert, Claudia Giesecke, Joachim R Grün, Tobias Alexander, Stefanie Schmidt, Katarzyna Luda, Anja A Kühl, Robby Engelmann, Michael Dürr, Tobias Scheel, Maria Bokarewa, Carsten Perka, Andreas Radbruch, and Thomas Dörner. A unique population of IgG-expressing plasma cells lacking CD19 is enriched in human bone marrow. *Blood*, 125(11):1739–48, March 2015.

[49] Thierry Mora, Aleksandra M Walczak, William Bialek, and Curtis G Callan. Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 107(12):5405–10, March 2010.

[50] E Paramithiotis and M D Cooper. Memory B lymphocytes migrate to bone marrow in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 94(1):208–12, January 1997.

[51] Andreas Radbruch, Gwendolin Muehlinghaus, Elke O Luger, Ayako Inamine, Kenneth G C Smith, Thomas Dörner, and Falk Hiepe. Competence and competition: the challenge of becoming a long-lived plasma cell. *Nature reviews. Immunology*, 6(10):741–50, October 2006.

[52] Grzegorz A Rempala, Micha Seweryn, and Leszek Ignatowicz. Model for comparative analysis of antigen receptor repertoires. *Journal of theoretical biology*, 269(1):1–15, January 2011.

[53] Florian Rubelt, Christopher R Bolen, Helen M McGuire, Jason A Vander Heiden, Daniel Gadala-Maria, Mikhail Levin, Ghia M Euskirchen, Murad R Mamedov, Gary E Swan, Cornelia L Dekker, Lindsay G Cowell, Steven H Kleinstein, and Mark M Davis. Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. *Nature communications*, 7:11112, March 2016.

[54] M K Slifka, R Antia, J K Whitmire, and R Ahmed. Humoral immunity due to long-lived plasma cells. *Immunity*, 8(3):363–72, March 1998.

[55] Reuben M Tooze. A replicative self-renewal model for long-lived plasma cells: questioning irreversible cell cycle exit. *Frontiers in immunology*, 4:460, December 2013.

[56] Johannes Trück, Maheshi N Ramasamy, Jacob D Galson, Richard Rance, Julian Parkhill, Gerton Lunter, Andrew J Pollard, and Dominic F Kelly. Identification of antigen-specific B cell receptor sequences using public repertoire analysis. *Journal of immunology (Baltimore, Md. : 1950)*, 194(1):252–61, January 2015.

[57] Renee C Tschumper, Yan W Asmann, Asif Hossain, Paul M Huddleston, Xiaosheng Wu, Angela Dispenzieri, Bruce W Eckloff, and Diane F Jelinek. Comprehensive assessment of potential multiple myeloma immunoglobulin heavy chain V-D-J intraclonal variation using massively parallel pyrosequencing. *Oncotarget*, 3(4):502–13, April 2012.

[58] Christophe Viret and Werner Gurr. The origin of the "one cell-one antibody" rule. *Journal of immunology (Baltimore, Md. : 1950)*, 182(3):1229–30, February 2009.

[59] R Wasserman, Y Ito, N Galili, M Yamada, B A Reichard, S Shane, B Lange, and G Rovera. The pattern of joining (JH) gene usage in the human IgH chain is established predominantly at the B precursor cell stage. *Journal of immunology (Baltimore, Md. : 1950)*, 149(2):511–6, July 1992.

[60] Joshua A Weinstein, Ning Jiang, Richard A White, Daniel S Fisher, and Stephen R Quake. High-throughput sequencing of the zebrafish antibody repertoire. *Science (New York, N.Y.)*, 324(5928):807–10, May 2009.

[61] Yael Weiss-Ottolenghi and Jonathan M Gershoni. Profiling the IgOme: Meeting the challenge. *FEBS letters*, 588(2):318–25, January 2014.

[62] Yariv Wine, Daniel R Boutz, Jason J Lavinder, Aleksandr E Miklos, Randall A Hughes, Kam Hon Hoi, Sang Taek Jung, Andrew P Horton, Ellen M Murrin, Andrew D Ellington, Edward M Marcotte, and George Georgiou. Molecular deconvolution of the monoclonal antibodies that comprise the polyclonal serum response. *Proceedings of the National Academy of Sciences of the United States of America*, 110(8):2993–8, February 2013.

[63] Jens Wrammert, Dimitrios Koutsonanos, Gui-Mei Li, Srilatha Edupuganti, Jianhua Sui, Michael Morrissey, Megan McCausland, Ioanna Skountzou, Mady Hornig, W Ian Lipkin, Aneesh Mehta, Behzad Razavi, Carlos Del Rio, Nai-Ying Zheng, Jane-Hwei Lee, Min Huang, Zahida Ali, Kaval Kaur, Sarah Andrews, Rama Rao Amara, Youliang Wang, Suman Ranjan Das, Christopher David O'Donnell, Jon W Yewdell, Kanta Subbarao, Wayne A Marasco, Mark J Mulligan, Richard Compans, Rafi Ahmed, and Patrick C Wilson. Broadly cross-reactive antibodies dominate the human B cell response against 2009 pandemic H1N1 influenza virus infection. *The Journal of experimental medicine*, 208(1):181–93, January 2011.

[64] Gabriel Wu, Nai-Kong Cheung, George Georgiou, Edward Marcotte, and Gregory C Ippolito. Temporal stability and molecular persistence of the bone marrow plasma cell antibody repertoire. *bioRxiv*, 2016.

[65] Xueling Wu, Tongqing Zhou, Jiang Zhu, Baoshan Zhang, Ivelin Georgiev, Charlene Wang, Xuejun Chen, Nancy S Longo, Mark Louder, Krisha McKee, Sijy O'Dell, Stephen Perfetto, Stephen D Schmidt, Wei Shi, Lan Wu, Yongping Yang, Zhi-Yong Yang, Zhongjia Yang, Zhenhai Zhang, Mattia Bonsignori, John A Crump, Saidi H Kapiga, Noel E Sam, Barton F Haynes, Melissa Simek, Dennis R Burton, Wayne C Koff, Nicole A Doria-Rose, Mark Connors, James C Mullikin, Gary J Nabel, Mario Roederer, Lawrence Shapiro, Peter D Kwong, and John R Mascola. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science (New York, N.Y.)*, 333(6049):1593–602, September 2011.

[66] J L Xu and M M Davis. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity*, 13(1):37–45, July 2000.

[67] Gur Yaari and Steven H Kleinstein. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome medicine*, 7:121, November 2015.

[68] Hao Zhang, Weidong Cui, and Michael L Gross. Mass spectrometry for the biophysical characterization of therapeutic monoclonal antibodies. *FEBS letters*, 588(2):308–17, January 2014.

[69] Rolf Zinkernagel. On plasma cell longevity or brevity. *Expert review of vaccines*, 13(7):821–3, July 2014.

Vita

Gabriel Chi Sun Wu was born in San Francisco, California and raised in Daly City. He graduated from Saint Ignatius College Preparatory. He then attended the University of California, Berkeley. While there, he did research on skin friction with Drs. Howard Maibach and Dorian Liepmann, and participated on Cal's inaugural International Genetically Engineered Machine (iGEM) Competition team. After receiving a Bachelor of Science degree in Bioengineering, Gabriel continued research in synthetic biology in the lab of Dr. John Dueber. He eventually entered doctoral studies at the University of Texas at Austin in Cell and Molecular Biology, where he was an NSF Graduate Research Fellow in the lab of Dr. Edward Marcotte. In 2014, he attended the 64th Lindau Nobel Laureate Meeting. In 2016, he chaired the Paul D. Gottlieb Endowed Lecture Series hosting Chemistry Nobel Laureate, W.E. Moerner. Gabriel has presented his scientific work internationally in Hong Kong and Rio De Janeiro. He is the author of ten peer-reviewed scientific studies and one patent. He has also played violin as part of the Villa Chamber Orchestra with performances in Germany, Austria, England, Ireland, and Carnegie Hall.

Permanent address: wug@utexas.edu

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.